A Companion to the Philosophy of Language

Blackwell Companions to Philosophy

This outstanding student reference series offers a comprehensive and authoritative survey of philosophy as a whole. Written by today's leading philosophers, each volume provides lucid and engaging coverage of the key figures, terms, topics, and problems of the field. Taken together, the volumes provide the ideal basis for course use, representing an unparalleled work of reference for students and specialists alike.

Already published in the series:

1. The Blackwell Companion to Philosophy, Second Edition
   Edited by Nicholas Bunnin and Eric Tsui-James
2. A Companion to Ethics
   Edited by Peter Singer
3. A Companion to Aesthetics, Second Edition
   Edited by Stephen Davies, Kathleen Marie Higgins, Robert Hopkins, Robert Stecker, and David E. Cooper
4. A Companion to Epistemology, Second Edition
   Edited by Jonathan Dancy, Ernest Sosa, and Matthias Steup
5. A Companion to Contemporary Political Philosophy (two-volume set), Second Edition
   Edited by Robert E. Goodin and Philip Pettit
6. A Companion to Philosophy of Mind
   Edited by Samuel Guttenplan
7. A Companion to Metaphysics, Second Edition
   Edited by Jaegwon Kim, Ernest Sosa, and Gary S. Rosenkrantz
8. A Companion to Philosophy of Law and Legal Theory, Second Edition
   Edited by Dennis Patterson
9. A Companion to Philosophy of Religion, Second Edition
   Edited by Charles Taliaferro, Paul Draper, and Philip L. Quinn
10. A Companion to the Philosophy of Language
    Edited by Bob Hale and Crispin Wright
11. A Companion to World Philosophies
    Edited by Eliot Deutsch and Ron Bontekoe
12. A Companion to Continental Philosophy
    Edited by Simon Critchley and William Schroeder
13. A Companion to Feminist Philosophy
    Edited by Alison M. Jaggar and Iris Marion Young
14. A Companion to Cognitive Science
    Edited by William Bechtel and George Graham
15. A Companion to Bioethics, Second Edition
    Edited by Helga Kuhse and Peter Singer
16. A Companion to the Philosophers
    Edited by Robert L. Arrington
17. A Companion to Business Ethics
    Edited by Robert E. Frederick
18. A Companion to the Philosophy of Science
    Edited by W. H. Newton-Smith
19. A Companion to Environmental Philosophy
    Edited by Dale Jamieson
20. A Companion to Analytic Philosophy
    Edited by A. P. Martinich and David Sosa
21. A Companion to Genethics
    Edited by Justine Burley and John Harris
22. A Companion to Philosophical Logic
    Edited by Dale Jacquette
23. A Companion to Early Modern Philosophy
    Edited by Steven Nadler
24. A Companion to Philosophy in the Middle Ages
    Edited by Jorge J. E. Gracia and Timothy B. Noone
25. A Companion to African-American Philosophy
    Edited by Tommy L. Lott and John P. Pittman
26. A Companion to Applied Ethics
    Edited by R. G. Frey and Christopher Heath Wellman
27. A Companion to the Philosophy of Education
    Edited by Randall Curren
28. A Companion to African Philosophy
    Edited by Kwasi Wiredu
29. A Companion to Heidegger
    Edited by Hubert L. Dreyfus and Mark A. Wrathall
30. A Companion to Rationalism
    Edited by Alan Nelson
31. A Companion to Pragmatism
    Edited by John R. Shook and Joseph Margolis
32. A Companion to Ancient Philosophy
    Edited by Mary Louise Gill and Pierre Pellegrin
33. A Companion to Nietzsche
    Edited by Keith Ansell Pearson
34. A Companion to Socrates
    Edited by Sara Ahbel-Rappe and Rachana Kamtekar
35. A Companion to Phenomenology and Existentialism
    Edited by Hubert L. Dreyfus and Mark A. Wrathall
36. A Companion to Kant
    Edited by Graham Bird
37. A Companion to Plato
    Edited by Hugh H. Benson
38. A Companion to Descartes
    Edited by Janet Broughton and John Carriero
39. A Companion to the Philosophy of Biology
    Edited by Sahotra Sarkar and Anya Plutynski
40. A Companion to Hume
    Edited by Elizabeth S. Radcliffe
41. A Companion to the Philosophy of History and Historiography
    Edited by Aviezer Tucker
42. A Companion to Aristotle
    Edited by Georgios Anagnostopoulos
43. A Companion to the Philosophy of Technology
    Edited by Jan-Kyrre Berg Olsen, Stig Andur Pedersen, and Vincent F. Hendricks
44. A Companion to Latin American Philosophy
    Edited by Susana Nuccetelli, Ofelia Schutte, and Otávio Bueno
45. A Companion to the Philosophy of Literature
    Edited by Garry L. Hagberg and Walter Jost
46. A Companion to the Philosophy of Action
    Edited by Timothy O'Connor and Constantine Sandis
47. A Companion to Relativism
    Edited by Steven D. Hales
48. A Companion to Hegel
    Edited by Stephen Houlgate and Michael Baur
49. A Companion to Schopenhauer
    Edited by Bart Vandenabeele
50. A Companion to Buddhist Philosophy
    Edited by Steven M. Emmanuel
51. A Companion to Foucault
    Edited by Christopher Falzon, Timothy O'Leary, and Jana Sawicki
52. A Companion to the Philosophy of Time
    Edited by Heather Dyke and Adrian Bardon
53. A Companion to Donald Davidson
    Edited by Ernest Lepore and Kirk Ludwig
54. A Companion to Rawls
    Edited by Jon Mandle and David Reidy
55. A Companion to W.V.O Quine
    Edited by Gilbert Harman and Ernest Lepore
56. A Companion to Derrida
    Edited by Zeynep Direk and Leonard Lawlor
57. A Companion to David Lewis
    Edited by Barry Loewer and Jonathan Schaffer
58. A Companion to Kierkegaard
    Edited by Jon Stewart
59. A Companion to Locke
    Edited by Matthew Stuart
60. The Blackwell Companion to Hermeneutics
    Edited by Niall Keane and Chris Lawn
61. A Companion to Ayn Rand
    Edited by Allan Gotthelf and Gregory Salmieri
62. The Blackwell Companion to Naturalism
    Edited by Kelly James Clark

Forthcoming:

A Companion to Mill
Edited by Christopher Macleod and Dale E. Miller

A COMPANION TO THE PHILOSOPHY OF LANGUAGE

SECOND EDITION

Volume I

Edited by

Bob Hale, Crispin Wright,
and Alexander Miller

WILEY Blackwell

A COMPANION TO THE PHILOSOPHY OF LANGUAGE

SECOND EDITION

Volume II

Edited by

Bob Hale, Crispin Wright,
and Alexander Miller

WILEY Blackwell

This second edition first published 2017

Edition history: Blackwell Publishing Ltd. (1e, 1997)

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication data is available for this title

Hardback ISBN: 9781118974711

A catalogue record for this book is available from the British Library.

Cover image: *The Chateau of Medan*, by Paul Cezanne (1839–1906) /
© CSG CIC Glasgow Museums and Libraries Collections

Set in 10/12.5pt Minion by SPi Global, Pondicherry, India

1 2017

Contents

VOLUME I

List of Contributors

Paul Artin Boghossian is Silver Professor of Philosophy at New York University and Director of its New York Institute of Philosophy. He has published many papers on the philosophy of mind and the philosophy of language, on such topics as color, rule-following, eliminativism, naturalism, self-knowledge, *a priori* knowledge, analytic truth, realism, and relativism. He is the author of *Fear of Knowledge* (Oxford University Press, 2006) and co-editor of *New Essays on the A Priori* (with Christopher Peacocke; Oxford University Press, 2000). A collection of his essays – *Content and Justification* – was published by Oxford University Press in 2008. A series of exchanges with Timothy Williamson, some previously published some new, on the analytic and the *a priori*, will appear from Oxford University Press in 2018.

Berit Brogaard is Director of the Brogaard Lab for Multisensory Research and Professor of Philosophy at the University of Miami. Her areas of research include perception, consciousness, emotions, philosophical psychology, semantics, and philosophical logic. She has written three books: *Transient Truths* (Oxford University Press, 2012), *On Romantic Love* (Oxford University Press, 2015), and *The Superhuman Mind* (Penguin, 2015), as well as over one hundred peer-reviewed articles.

John Collins is Professor of Philosophy at the University of East Anglia. He has published widely in the philosophy of language and mind, with especial reference to generative linguistics, and on the concept of truth. He is the author of *Chomsky: A Guide for the Perplexed* (Continuum, 2008) and *The Unity of Linguistic Meaning* (Oxford University Press, 2011), and co-editor of *Experimental Philosophy, Rationalism, and Naturalism* (with Eugen Fischer; Routledge, 2015).

Edward Craig is former Knightbridge Professor of Philosophy at the University of Cambridge, and has been a Fellow of the British Academy since 1993. He is the author of *The Mind of God and the Works of Man* (Oxford University Press, 1987) and *Knowledge and the State of Nature* (Oxford University Press, 1990), as well as articles on various topics in the theory of knowledge and philosophy of language. He is chief editor of the *Routledge Encyclopedia of Philosophy*.

Andy Egan is Professor of Philosophy at Rutgers University. He has held positions at the University of Michigan and the Australian National University. He attended graduate school at the University of Colorado and Massachusetts Institute of Technology. He works primarily in philosophy of language and philosophy of mind.

Kit Fine is University Professor and Silver Professor of Philosophy and Mathematics at New York University. His areas of interest include philosophical logic, philosophy of language, and metaphysics and his more recent books include *Modality and Tense* (Oxford University Press, 2005) and *Semantic Relationism* (Blackwell, 2007). He is a Fellow of the American Academy of Arts and Letters, and a corresponding Fellow of the British Academy.

Graeme Forbes is Professor of Philosophy at the University of Colorado at Boulder. He is the author of *Attitude Problems* (Oxford University Press, 2006) and the textbook *Modern Logic* (Oxford University Press, 1994). He works mainly in semantics, metaphysics, and logic, and has interests in compositionality, intensionality, modal metaphysics, and modal logic.

Anthony S. Gillies is Professor of Philosophy at Rutgers University, and previously taught at the University of Michigan, Harvard, and the University of Texas at Austin, and was White Distinguished Visiting Professor at the University of Chicago. His research interests are in philosophy of language: formal semantics and pragmatics; epistemology: belief revision, defeasible reasoning; philosophical logic; and decision/game theory.

Bob Hale is an Emeritus Professor at the University of Sheffield, and his main research interests are in the foundations of mathematics, and philosophy of logic and language. He is a member of the editorial board of *Philosophia Mathematica*, and is author of *Abstract Objects* (Blackwell, 1987) and *Necessary Beings* (Oxford University Press, 2013; revised 2nd edn, 2015); co-editor of *Reading Putnam* (with Peter Clark; Blackwell, 1994); co-editor of *Modality: Metaphysics, Logic, and Epistemology* (with Aviv Hoffmann; Oxford University Press, 2010); and co-author of *The Reason's Proper Study* (with Crispin Wright; Oxford University Press, 2001).

Anandi Hattiangadi has been Professor of Philosophy at Stockholm University and Pro Futura Scientia Fellow at the Swedish Collegium of Advanced Studies since 2013, before which she was a tutorial fellow of St Hilda's College, Oxford. She has research interests in the philosophy of mind and language, epistemology, metaphysics, and metaethics. Her publications include *Oughts and Thoughts: Rule-Following and the Normativity of Content* (Oxford University Press, 2007), as well as numerous articles on the normativity of meaning, content, and belief.

Jussi Haukioja is Professor of Philosophy at the Norwegian University of Science and Technology and editor of the volume *Advances in Experimental Philosophy of Language* (Bloomsbury, 2015). His research interests are in philosophy of language, philosophy of mind, and realism and anti-realism.

Jane Heal is Emeritus Professor of Philosophy at the University of Cambridge and a Fellow of St John's College. Her interests are mainly in philosophy of language and

philosophy of mind. Her previous publications include her book *Fact and Meaning* (Blackwell, 1989) and several journal articles in these areas. She was elected a Fellow of the British Academy in 1997.

Barry Loewer is Professor and Director of the Rutgers Center for Philosophy and the Sciences. His published work lies mainly in the philosophy of mind and psychology, the philosophy of quantum mechanics, and metaphysics, including the book *Why There is Anything Except Physics* (Oxford University Press, 2008).

Guy Longworth is Associate Professor in Philosophy at the University of Warwick. He works mainly in the philosophy of language and mind, including intersections with epistemology.

E. J. Lowe was Professor of Philosophy at the University of Durham, where he taught from 1980 until his death in 2014. He authored 11 books, including *Kinds of Being* (Blackwell, 1989) and *The Possibility of Metaphysics* (Oxford University Press, 1998), and also co-edited four volumes and wrote over two hundred articles for journals and edited collections.

Michael P. Lynch is Professor of Philosophy and Director of the Humanities Institute at the University of Connecticut. He is the author or editor of seven books including *In Praise of Reason* (MIT Press, 2012), *Truth as One and Many* (Oxford University Press, 2009), and *True to Life* (MIT Press, 2004). His research interests lie in pursuing problems within the intersection of epistemology, metaphysics, and the philosophy of language.

Penelope Mackie is Associate Professor and Reader in Philosophy at the University of Nottingham. She is the author of *How Things Might Have Been: Individuals, Kinds, and Essential Properties* (Oxford University Press, 2006) and of a number of articles on topics in metaphysics, including causation, modality, material constitution, free will, and the fixity of the past.

Aidan McGlynn is a lecturer in Philosophy at the University of Edinburgh. He recently completed a series of papers and a monograph on knowledge first approaches to epistemology and the philosophies of language and mind. Since then, he has been working on evidence, first-person thought and self-knowledge, pornography, epistemic injustice, silencing, and objectification.

Andrew McGonigal holds a visiting professorship in philosophy at Washington and Lee University. Before taking up the position, he taught for 12 years at the University of Leeds. He is a co-editor of the *Routledge Companion to Metaphysics*, and in 2014–2015 was awarded a Society Fellowship at the Society for the Humanities at Cornell.

Alexander Miller is Professor of Philosophy and chair of the Department of Philosophy at the University of Otago. He works mainly on the philosophy of language and mind, metaphysics, and metaethics. His books include *Contemporary Metaethics: An Introduction Revised and Expanded* (2nd edn, Polity Press, 2013) and *Philosophy of Language Revised and Expanded* (2nd edn, Routledge, 2007). He is co-editor of *Rule-Following and Meaning* (with Crispin Wright; Acumen, 2002).

Richard Moran is Professor of Philosophy at Harvard University, having previously taught at Princeton University. He works primarily in the areas of moral psychology, the philosophy of mind and language, aesthetics and the philosophy of literature, and the later Wittgenstein. He has published papers on metaphor, on imagination and emotional engagement with art, and on the nature of self-knowledge. His book, *Authority and Estrangement: An Essay on Self-Knowledge*, was published by Princeton University Press in 2001.

Michael Morris is Professor of Philosophy at the University of Sussex. He is the author of *The Good and the True* (Oxford University Press, 1992), *An Introduction to the Philosophy of Language* (Cambridge University Press, 2007), and *Wittgenstein and The Tractatus* (Routledge, 2008), as well as papers in the philosophy of language and the philosophy of art.

Julien Murzi completed his PhD at the University of Sheffield in October 2010. He is Assistant Professor at the University of Salzburg, having previously been a post-doctoral fellow at the Munich Center for Mathematical Philosophy (of which he continues to be an external member) and Lecturer in Philosophy at the University of Kent. He has published papers on inferentialism, logical consequence, the semantic paradoxes, the realism/anti-realism debate, and the open future.

Bernhard Nickel is Professor of Philosophy at Harvard University. He works mainly in philosophy of language and semantics, with interests in metaphysics, the philosophy of science, and philosophy of mind. He is the author of *Between Logic and the World* (Oxford University Press, 2016), which presents a theory of generics and genericity.

Christian Nimtz is Professor of Theoretical Philosophy at Bielefeld University in Germany. His interests lie mainly in the philosophy of language, modal epistemology, and meta-philosophy. He has worked on natural kind terms, modal knowledge, thought experiments, and conceptual analysis.

Harold Noonan is Professor of Mind and Cognition at the University of Nottingham. He has published seven books, including *Hume* (One Word Publishers, 2007) and *Frege* (Polity Press, 2001), as well as various articles on topics in the philosophy of mind, philosophy of language, and philosophy of logic.

Christopher Peacocke is Johnsonian Professor of Philosophy at Columbia University, and was previously Waynflete Professor of Metaphysical Philosophy at the University of Oxford, where he also held a Leverhulme Personal Research Professorship. He is the author of several books, most recently *The Mirror of the World: Subjects, Consciousness, and Self-Consciousness* (Oxford University Press, 2014), and of papers in the philosophy of language, mind, psychology, and logic. He is a Fellow of the British Academy and the American Academy of Arts and Sciences.

John Perry is the Waldgrave Stuart Professor of Philosophy Emeritus at Stanford University and Distinguished Professor of Philosophy Emeritus at the University of California, Riverside. He is co-director of the Center for the Explanation of Consciousness at the Center for the Study of Language and Information. He has authored several books, including the second

enlarged edition of *Reference and Reflexivity* (CSLI Publications, 2012) and various articles on the philosophy of language. He also co-hosts a weekly talk show called *Philosophy Talk*.

James Pryor is a Professor of Philosophy at New York University. His research focus is epistemology, formal semantics (especially issues at the intersection of philosophy, linguistics, and computer science), philosophy of mind, and related issues.

Mark Richard is Professor of Philosophy at Harvard University. He works in the philosophy of language, epistemology, and metaphysics, as well as in mathematical and intensional logic, philosophy of logic, and philosophy of mind. He owns a Fender Stratocaster but sadly at the moment lacks a dog. He is the author of numerous articles and books, most recently *Meaning in Context, Volume I: Context and the Attitudes* (Oxford University Press, 2013) and *Meaning in Context, Volume II: Truth and Truth Bearers* (Oxford University Press, 2015).

Ian Rumfitt is a Senior Research Fellow of All Souls College, Oxford. He works mainly in the philosophy of language, philosophical logic, and the philosophy of mathematics. His book *The Boundary Stones of Thought* (Oxford University Press, 2015) investigates conflicts between rival logical systems and how they might be rationally resolved.

Mark Sainsbury is Professor of Philosophy at the University of Texas at Austin, having formerly taught at King's College London. He is the author of *Russell* (Routledge, 1979), *Paradoxes* (Cambridge University Press, 1987), *Logical Forms* (Blackwell, 1991), *Departing From Frege* (Routledge, 2002), *Reference Without Referents* (Oxford University Press, 2005), and *Fiction and Fictionalism* (Routledge, 2009), and co-author of *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts* (with Michael Tye; Oxford University Press, 2013). His *Thinking About Things* is due out from Oxford University Press in 2017.

Stephen Schiffer is Silver Professor of Philosophy at New York University. He works primarily in philosophy of language, philosophy of mind, and metaphysics. He is the author of numerous articles and of three books: *Meaning* (Oxford University Press, 1972), *Remnants of Meaning* (MIT Press, 1987), and *The Things We Mean* (Oxford University Press, 2003). He is a Fellow of the American Academy of Arts and Sciences.

Peter Schulte teaches philosophy at the Bielefeld University. His areas of specialization are philosophy of mind, metaphysics, philosophy of language, metaethics, and free will.

Patrick Shirreff received a BA (Hons) from the University of Toronto in 2010 and is an ABD at the University of Michigan. His research focuses on the intersection of philosophy of language and epistemology. Specifically, Shirreff is interested in the semantics of epistemic language and what this semantic theorizing can show us about epistemic theorizing.

John Skorupski is Professor Emeritus of Moral Philosophy at the University of St Andrews. His current interests are in moral and political philosophy, metaethics and epistemology, and the history of nineteenth- and twentieth-century philosophy. His most recent books are *The Domain of Reasons* (Oxford University Press, 2010) and *Why Read Mill Today?* (Routledge, 2006).

Robert Stalnaker is Professor of Philosophy in the Department of Linguistics and Philosophy at MIT. His teaching and research interests are in philosophical logic, philosophy of mind, and the philosophy of language. He is the author of *Inquiry* (MIT Press, 1984), and of various articles on intentionality and the foundations of semantics and pragmatics. He also has two volumes of collected papers: *Context and Content* (Oxford University Press, 1999) and *Ways a World Might Be* (Oxford University Press, 2003).

Jason Stanley is Jacob Urowsky Professor of Philosophy at Yale University. His interests include the philosophy of language, the history and philosophy of logic, the history of analytic philosophy, epistemology, and the philosophy of mind. He is the author of four books, most recently *How Propaganda Works* (Princeton University Press, 2015) and *Know How* (Oxford University Press, 2011).

Florian Steinberger joined the Department of Philosophy at the Birkbeck University of London in 2015, prior to which he was Assistant Professor in Philosophy and Language at the Ludwig-Maximilians University in Munich and the Munich Center for Mathematical Philosophy. His main research interests include parts of epistemology, normativity, and the philosophies of logic and language.

Charles Travis is Professor Emeritus at King's College London and a Professor Afiliado in the Faculdade de Letras at the University of Porto. He has published extensively on the philosophy of language and the philosophy of mind. He is the author of many books, including *Perception: Essays After Frege* (Oxford University Press, 2013) and *Objectivity and the Parochial* (Oxford University Press, 2011), together with numerous articles.

Ralph C. S. Walker is Emeritus Fellow at Magdalen College, Oxford. His research interests are in Immanuel Kant, ethics, philosophy of religion, truth, and justification of beliefs. His publications include *Kant* (Oxford University Press, 1978) and *The Coherence Theory of Truth* (Routledge, 1988).

Brian Weatherson is the Marshall M. Weinberg Professor of Philosophy at the University of Michigan, Ann Arbor. He works on epistemology, especially on issues at the intersection of ethics and epistemology, and issues at the interface between formal and traditional approaches to epistemology, as well as on many topics in philosophy of language.

Daniel Wee teaches philosophy at the Universiti Brunei Darussalam, having completed his PhD on rule-following and communitarianism at the University of Otago in 2016. His research interests are in philosophy of language, ethics, meta-philosophy, and critical thinking.

Bernhard Weiss is Professor of Philosophy at the University of Cape Town. He is the editor of the collection *Dummett on Analytical Philosophy* (Palgrave Macmillan, 2015), and author of two books: *Michael Dummett* (Acumen, 2002) and *How To Understand Language* (Acumen, 2010). His areas of interest concern philosophies of language, logic and mathematics, and realism and anti-realism.

David Wiggins is Professor Emeritus of Philosophy at Oxford. He was previously Professor of Philosophy at Birkbeck College, London, and before that Fellow and Praelector of University College, Oxford. His principal publications are *Sameness and Substance* (Blackwell, 1980) and *Needs, Values, Truth* (2nd edn, Blackwell, 1998), as well as *Ethics: Twelve Lectures on the Philosophy of Morality* (Cambridge University Press, 2006) and *Sameness and Substance Renewed* (Cambridge University Press, 2001). He is a Fellow of the British Academy.

Timothy Williamson has been Wykeham Professor of Logic at Oxford since 2000, and previously taught logic and metaphysics at the University of Edinburgh. His main research interests are in philosophical logic, epistemology, metaphysics, and philosophy of language. He is the author of *Identity and Discrimination* (Blackwell, 1990), *Vagueness* (Blackwell, 1994), *Knowledge and its Limits* (Oxford University Press, 2000), and most recently *Tetralogue* (Oxford University Press, 2015), as well as articles in journals of philosophy and logic.

Crispin Wright is Professor of Philosophy at New York University and Professor of Philosophical Research at the University of Stirling. His books include *Wittgenstein on the Foundations of Mathematics* (Harvard University Press, 1980), *Frege's Conception of Numbers as Objects* (Humanities Press, 1983), *Truth and Objectivity* (Harvard University Press,1992), *Realism, Meaning and Truth* (2nd edn, Blackwell, 1993), *The Blackwell Companion to Philosophy of Language* (with Bob Hale; Blackwell, 1997), *The Reason's Proper Study* (with Bob Hale; Clarendon Press, 2001), *Rails to Infinity* (Harvard University Press, 2001), and *Saving the Differences* (Harvard University Press, 2003). Two collections of his papers, *The Riddle of Vagueness* and *Imploding the Demon*, are currently in preparation.

Preface to the Second Edition

We have taken advantage of Wiley-Blackwell's generous offer to publish a second and significantly expanded version of the first (1997) edition of the *Companion* to update the original chapters and to publish a range of new chapters that both broaden and deepen the coverage provided in the earlier edition.

Of the 25 chapters in the first edition, 21 have been updated, either by the original author or by a new author specifically commissioned for that purpose. Many updates take the form of postscripts to the originals, although a few simply revise and update the text from the first edition. The first edition chapter on intention and convention has been replaced by an entirely new chapter on the topic by Stephen Schiffer. The only first edition chapters reprinted unchanged are those by Christopher Peacocke, Robert Stalnaker, and Jason Stanley.

In addition to the 21 updates to the first edition and Schiffer's new chapter on intention and convention, there are 16 wholly new chapters covering both foundational issues and issues relating to specific linguistic phenomena. We have retained the tripartite structure of the original and have added a few new entries to the glossary.

We're grateful to all of our authors, both old and new, for their excellent chapters, updates, and glossary entries, and to Mark Cooper and Allison Koska at Wiley-Blackwell for their support and patience. Thanks, too, to Marielle Suba for her work on formatting final versions of the chapters, and to Marguerite Nesling and Giles Flitney for assistance with copy-editing and proofreading.

Bob Hale, Alex Miller, and Crispin Wright

Preface to the First Edition

The recent proliferation of dictionaries and encyclopedias of philosophy has resulted in no shortage of companionship for the philosophical tourist whose desire is merely for a short excursion. Our *Companion* is intended as a guide for a more determined and ambitious explorer. Thus this is no alphabetized compendium of brief statements of the principal theoretical positions, concepts, and protagonists in recent and contemporary philosophy of language, but comprises, rather, 25 extended essays on a nucleus of the most central issues in the field, each of which has seen and continues to see important work.

All of our contributors are active in research on their selected topics. Each was invited to contribute a chapter somewhat along the lines of the *State of the Art* series which *Mind* initiated in the mid-1980s: a survey and analysis of recent trends in work on the topic in question, offering a bibliography of the more important literature and incorporating a substantial research component. Accordingly, these are chapters for a philosophically experienced – advanced undergraduate, graduate, or professional – readership. Each chapter is, however, written so as to presuppose a minimum of prior knowledge of its specific subject-matter, and so offers both a self-contained overview of the relevant issues and of the shape of recent discussion of them and, for readers who want it, an up-to-date preparation for extended study of the topic concerned. There are, naturally, numerous points of connection among the chapters, some of which will be obvious enough from their titles or from a quick glance at their opening sections; others have been indicated by explicit cross-referencing. We have attempted, in the glossary, to provide concise explanations of all of the more important technical or semi-technical terms actually employed in the various chapters, and of a good number of other terms of art which, though not actually used by any of our contributors, figure centrally in other published work on the issues. The result, as we hope, is an anthology which will both stimulate research in the philosophy of language and provide an up-to-date textbook for its advanced teaching for many years to come.

Few would now subscribe to the idea which prevailed for a while in some Anglo-American philosophical circles during the 1970s, that the philosophy of language is First Philosophy, and that great issues in, for instance, metaphysics, epistemology, and the philosophy of mind, are to be resolved by, in effect, recasting them as matters for treatment within the theory of meaning. But there is no doubt that philosophy of language continues to occupy a position of central importance in contemporary philosophy, nor that some of

the best and most influential philosophical writing of the latter half of this century, by some of the foremost philosophical thinkers of our time, has been accomplished in this area. The threefold division into which we have organized the chapters closely reflects the landscaping which these leading authors have given to the subject. Part I, on *Meaning and Theories of Meaning,* comprises chapters which are all concerned, in one way or another, with issues connected to the nature of language mastery that have loomed large in the writings of Davidson, Dummett, and Grice. Part II, on *Language, Truth, and Reality*, pivots around more metaphysical issues to do with meaning: with the ongoing debate about meaning-skepticism that has drawn on the writings of Kripke, Putnam, Quine, and Wittgenstein, and with the connections between issues to do with meaning and the various debates about realism, whose excavation has been led by Dummett. Finally, Part III, on *Reference, Identity, and Necessity*, focuses on issues which take center stage in – or at least, loom large in the stage-setting for – Kripke's *Naming and Necessity*. Together, the three parts cover almost every topic that anyone familiar with contemporary work in the philosophy of language would expect to receive extensive discussion in a volume of this kind. There are nevertheless some vacancies which we would have liked, ideally, to have filled. There is, for example, no chapter focusing on the concept of a *criterion* which the first generation of commentary elicited from Wittgenstein's *Philosophical Investigations*, nor – perhaps more grievous – did we succeed in the end in commissioning a suitable study of semantic externalism or of notions of supervenience.

It remains to express our gratitude to our contributors, both for their patience with our editorial suggestions and for the excellence of their contributions and valuable assistance with glossary entries; to our publishers for bearing with us while we put together a volume which has been inevitably subject to many delays; to the secretarial staff of the Philosophy Departments of the Universities of St Andrews and Glasgow for assistance with the preparation and standardization of typescripts; and to each other.

Bob Hale and Crispin Wright

PART I

Meaning and Theories of Meaning

1

Metaphysics, Philosophy, and the Philosophy of Language

MICHAEL MORRIS

1 Two Positions

Michael Dummett famously declared (Dummett, 1993, p. 4):

> What distinguishes analytical philosophy, in its diverse manifestations, from other schools is the belief, first, that a philosophical account of thought can be attained through a philosophical account of language, and, secondly, that a comprehensive account can only be so attained.

He had earlier claimed (Dummett, 1978e, p. 458):

> Only with Frege was the proper object of philosophy finally established: namely, first, that the goal of philosophy is the analysis of the structure of *thought*; secondly, that the study of *thought* is to be sharply distinguished from the study of the psychological process of *thinking*; and, finally, that the only proper method for analysing thought consists in the analysis of *language*.

In sharp contradistinction, Timothy Williamson, taking metaphysics to be "central" to philosophy (a point to note before moving on), asserts (Williamson, 2007, pp. 18–19):

> Much contemporary metaphysics is not primarily concerned with thought or language at all. Its goal is to discover what fundamental kinds of thing there are and what properties and relations they have, not to study the structure of our thought about them – perhaps we have no thought about them until it is initiated by metaphysicians. Contemporary metaphysics studies substances and essences, universals and particulars, space and time, possibility and necessity. Although nominalist or conceptualist reductions of all these matters have been attempted, such theories have no methodological priority and generally turn out to do scant justice to what they attempt to reduce.

We seem to have here the following stark contrast: Dummett thinks understanding language is central to philosophy, whereas Williamson apparently does not. Dummett is

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

endorsing some form of what has been known as the "linguistic turn"[1] – the dominant tendency in English-speaking philosophy in the middle of the twentieth century – whereas Williamson is rejecting it.

I offer here a selective critical history in which I trace the difference between the tendency which Dummett represents and the philosophers among whom Williamson is naturally placed to a difference in metaphysics which has much longer roots.[2] In fact it turns out that those who reject the tendency Dummett represents also often give a central role to the philosophy of language. This is questionable too, though on other grounds.

2 Dummett and Thought

I will not dwell on the fact that Dummett counts it a distinctive mark of *analytic* philosophy in particular to give priority to the philosophy of language. (I take it that he is here aiming to contrast analytic philosophy both with the philosophy which preceded it and against which it was a reaction (most obviously Hegelianism in various forms), and with the older philosophers in the phenomenological tradition (who may be said to give priority instead to a proper attentiveness to the actual character of experience).) Nor will I linger over the fact that his characterization of analytic philosophy is odd from a classificatory point of view, since it both excludes some of the most prominent analytic philosophers – Dummett himself acknowledges that Gareth Evans is left out, only counting as analytic "as belonging to this tradition" (Dummett, 1993, p. 4) – and includes some philosophers it would be odd to call analytic (Derrida is the obvious example here). I will simply take him here to be declaring in another way his view of how philosophy *ought* to be done.

When we set that issue aside, what is most immediately striking about what Dummett says is not the importance he gives to the philosophy of language, but the importance he gives to providing a philosophical account of *thought*. It seems more natural to think that the business of philosophy is to make sense, in the first instance, not of thought, but of the *world* – which is to say, of the *objects* of thought. Of course thought itself may be thought about, and so itself be an object of thought, but it is natural to expect philosophy to be concerned with thought chiefly when it is the object of thought, which is not all that often. Why, then, does Dummett give such central importance to the task of making sense of thought?

His official reason appears to be that "Thoughts [in the sense of *what is thought*] differ from all else that is said to be among the contents of the mind in being wholly communicable" (Dummett, 1978e, p. 442). But we have seen that Dummett takes recognizing the importance of the philosophy of language to depend on the antecedent recognition of the importance of an understanding of thought, whereas this remark makes thought important only in so far as we are interested in communication – an interest which looks as if it depends on an interest in language.

We might suggest that the importance Dummett gives to thought depends on his interest in Frege.[3] In Frege's later philosophy, thoughts – understood as the senses of sentences, what are expressed by sentences – might be taken to be the principal focus of his concerns; and thoughts are certainly the primary bearers of truth for Frege (Frege, 1977). But again it is hard to see how this can be the ultimate explanation of Dummett's focus on thought. First, Frege seems to be concerned with thought only because he is already concerned with

language, so, as before, this fails to explain why Dummett should want to explain the importance of the philosophy of language in terms of the importance of understanding thought. And second, the first and most striking case where Frege seems to give language a central importance is a case in which a claim about language is used directly to make a claim about the nature of the *world*, without any detour through thought. One of Frege's crucial claims in the *Foundations of Arithmetic* is that numbers are objects (Frege, 1953). It is already clear that Frege takes numbers to belong to the world – to what would later be counted as the realm of reference – rather than to anything psychological (and at this point he had not isolated a distinct realm of sense). The claim that they are objects depends just on two further claims: first, that objects are nothing but the referents of singular terms; and, second, that number words are best understood as singular terms. This is a case of a philosophical account of language being taken to be the only way of achieving a philosophical account of the *world*, not of thought.

So if we are to make sense of Dummett's giving such central importance to the task of making sense of thought we need to look elsewhere. I think the place to look is obvious enough, when we think of Dummett's links to verificationism and to other strands of the empiricist tradition.[4] I suggest that the ultimate source of the kind of role Dummett gives to thought is Hume's skeptical view of necessity, with its famous consequences for metaphysics.[5]

Hume's view depends on empiricism about our grip on reality, combined with a particular theory of perception which he shared, in general outline at least, with the other classical empiricists, and with most other philosophers down to the middle of the twentieth century. His general empiricism requires that if we are to have any knowledge of something in the world – something which is, in some sense, independent of us – it must be made available to us through sense-perception. (This also, of course, limits what can be included in the world – in what is truly independent of us – at least in so far as we can have knowledge of it.) So if necessity and possibility (the necessity and possibility which concern us, at least) are to be in the world – to be, in the relevant sense, independent of us – they have to be made available to us through perception. And conversely, if necessity and possibility are not made available to us through perception, then when we take ourselves to be thinking of necessity and possibility we cannot be thinking of something which is properly independent of us. The theory of perception adds further constraints to this general empiricist picture in two stages. It first limits what can genuinely be perceived to what can be constructed from what is distinctively available to each of the senses (color to sight, sound to hearing, and so on). It then limits what can genuinely be perceived still further by insisting that what is genuinely or immediately perceived must be constant between genuine, veridical perception, on the one hand, and illusion or hallucination, on the other (so if something could have looked the same even if it hadn't had some feature, then it is impossible genuinely or immediately to perceive that it has that feature). The theory of perception makes it hard to believe that necessity and possibility can be made available to us through perception, and the general empiricism then means that what we think of when we take ourselves to be thinking of necessity and possibility cannot be independent of us in the way which is required for them to be real, or really part of the world. The conclusion Hume draws seems inescapable (Hume, 1978, p. 165):

> Upon the whole, necessity is something, that exists in the mind, not in objects; nor is it possible for us ever to form the most distant idea of it, consider'd as a quality in bodies.

As this brief quotation makes clear, Hume contrasts what is genuinely in the world, in some sense independent of us – in this case as a 'quality in bodies' – with what depends on us. (There may be a question whether this contrast can be maintained consistently with all of the rest of his philosophy, but we do not need to pursue that question here.) If realism about something is the view that its nature is independent of us and of the way we think about it or represent it, and the correlative anti-realism is the rejection or non-acceptance of realism, Hume is naturally seen as favoring a general realism about the world – at least relatively speaking[6] – while adopting an anti-realist view about modality in particular: the world is real and independent of us, but possibility and necessity are not strictly part of the world.

This contrast, in turn, looks as if it forces us to accept a sharp division between kinds of discipline or enquiry, if we want to allow that there are any necessary truths at all. On the one hand, there are those disciplines or investigations which provide knowledge about the world. On the other hand, there are those disciplines or investigations which enable us to draw conclusions of necessity and possibility. The Humean combination of realism about the world and anti-realism about modality requires that these two kinds of discipline are fundamentally distinct: in so far as some discipline enables us to draw modal conclusions, it cannot be concerned with the real world; and in so far as it provides knowledge of the real world, its conclusions cannot be modal. This is, in effect, Hume's Fork.[7]

This sharp distinction looks as if it causes deep problems for metaphysics, on a natural understanding of metaphysics. On that natural understanding, the central business of metaphysics is to discover how the world must be – to discover its necessary features (see, e.g., Kant, 1997, B19–24, and Williamson, 2007, p. 134). On this conception, it is essential to metaphysics that it provide us with knowledge of truths which are both genuinely about the world and necessary. Hume's view then makes metaphysics impossible, as he noted with some glee (Hume, 1975, p. 165). Kant recognized this too, and attempted a defense of metaphysics in the face of the threat of Humeanism (Kant, 1997, B19–20). The Humean problem arises from the contrast between realism about the world and anti-realism about modality. Kant's solution – at least, on an orthodox interpretation – is to remove the contrast by weakening the realism about the world. The essence of Hume's view of necessity remains in place: it is just that the world, while being allowed to be real enough for everyday concerns, is no more real – no more independent of us – than necessity and possibility.

Hume seems to have to think that modal conclusions can only be based on reasoning concerning "abstract relations of our ideas" (Hume, 1978, p. 413), and ideas, for Hume, are components of thought. Kant, similarly – on the received interpretation I am following – seems to have taken the conception of the world in which modality has its home to be derived from judgment, whose structure and character is then inevitably reflected in the world which we think about. And this gives us reason to give thought a special place in philosophy, in so far as we think that philosophical conclusions are modal. If we follow the Humean view, and restrict the range of modal truths in his way, philosophy can only be about or expressive of thought and conceptual relations, and cannot reveal the nature of the world. If we follow Kant, and hope for a more ambitious metaphysics, we may be able to acquire knowledge of necessary truths about the world through philosophy, but in so far as the truths are necessary, they will reflect something in the structure of thought.

This conception of things is reflected in the key terms Kant used, which were to shape views of the nature of philosophy up until the latter part of the twentieth century. Crucial here are the terms 'analytic' and 'synthetic.' The class of analytic truths coincides roughly with what Hume would count as truths concerning 'abstract relations of our ideas.' And the

class of synthetic truths coincides roughly with truths which can be said to be genuinely revealing of the world. On Kant's account, necessary truths have to be *a priori* (Kant, 1997, B3), so in order to make sense of the possibility of metaphysics, we have to make sense of the possibility of synthetic *a priori* truths (Kant, 1997, A9–10; B13–14). Because he insists that what we need for metaphysics is something which is at least *a priori*, and he understands the *a priori* as what does not depend on sense-experience, Kant is in effect accepting the Humean view of the importance of the division between what does and what does not depend on sense-experience. In effect, he accepts the empiricist view that the world – at least as we can have knowledge of it – will be what we can have sense-experience of. Since Kant takes the *a priori* to be, roughly speaking, what we bring to experience, rather than what we get from it, allowing that there can be synthetic *a priori* truths already seems to commit him to the view that the world as we can know it depends on what we bring to experience.

If we interpret Dummett's conception of philosophy against this background, the importance he gives to the philosophy of language seems ultimately to depend on an antecedent commitment to the importance of metaphysics (or the nearest we can get to it) within philosophy as a whole, even if it goes by way of a form of anti-realism about modality. We can offer the following roughly formulated argument on his behalf:

(1) Metaphysics (or the nearest we can get to it) is the most fundamental philosophy;
(2) Metaphysics (or the nearest we can get to it) is to be pursued through an understanding of thought;
(3) Thought is to be understood through the philosophy of language; *so*
(4) Metaphysics (or the nearest we can get to it) is to be pursued through the philosophy of language; *and*
(5) The philosophy of language is the key to the most fundamental philosophy.

We were struck earlier by the importance Dummett gives to an understanding of thought. That is expressed here in (2), which gives voice to a form of anti-realism about modality. But we can get a similar general view about the importance of the philosophy of language within philosophy as a whole without talking much about thought at all. We get a more direct argument for the same conclusion if we simply omit (2) and (3), taking (5) to follow directly (as it does) from (1) and (4). I suggest that this second, more direct argument expresses a long-standing and still widely prevailing view. This more direct argument is itself, strictly speaking, neutral on the question of realism about modality.[8] It has certainly been understood through the lens of a broadly Humean anti-realism about modality, and I will look first at versions of the view which might seem to depend on that kind of understanding: these were dominant in the English-speaking world in the middle third of the twentieth century. But I will then turn to more recent approaches to philosophy, which also give a central role to the philosophy of language; these can also be understood as adhering to something like the more direct argument of (1), (4), and (5), and they are at least compatible with some form of realism about modality.

3 Wittgenstein, Early and Late

If you adopt a Humean combination of general realism about the world and anti-realism about modality, while thinking of metaphysics as the discovery of necessary truths about the world, you are likely to count metaphysics as being of relatively slight importance. If you

also hold (1), that metaphysics (or the nearest we can get to it) is the most fundamental philosophy, you will count philosophy in general as being of slight importance. This kind of view is associated with Ludwig Wittgenstein, in both his earlier and his later philosophy. And in both cases the approach to philosophy comes with the view that the most fundamental philosophy is to be approached through an understanding of language. I will suggest that although the early philosophy was one of the direct inspirations for a neo-Humean movement in philosophy – known variously as logical empiricism or logical positivism – it is in fact more Kantian than Humean. The later philosophy, however, can be seen as more simply Humean. I will begin with the later work, because this will enable us to understand more clearly how Wittgenstein's work as a whole connects with the Humean tradition.

Here is one of the most simply revealing sequences in Wittgenstein's later philosophy (2009b, 315):

> I can know what someone else is thinking, not what I am thinking.
>
> It is correct to say "I know what you are thinking," and wrong to say "I know what I am thinking."
>
> (A whole cloud of philosophy condenses into a drop of grammar.)

The first sentence here makes a modal claim: it asserts that something is *possible* (knowing what someone else is thinking) and that something else is *impossible* (knowing what one is thinking oneself). This seems to be exactly the kind of modal claim which Hume thought was spurious: it appears to say something about what is possible and what impossible *in the world* (some things really *can* be known, others really *cannot*).[9]

The second sentence then offers a parallel claim about what it is correct or incorrect to *say* – a claim about the rules for the use of the word 'know.' And the third sentence, in effect, claims that the sentence about linguistic rules gives the essence of the apparently modal claim of the first sentence. A way of putting the point here would be to say that what seems to be a modal claim about the world is really no more than an expression of a truth about the rules for using a word. This is naturally understood as a form of *projectivism* about necessity. Projectivism about a given subject-matter is the view that what we take to be real features of the world are in fact just projections of features of our ways of thinking of or representing the world. It gets its classic statement in this sentence of Hume's (where he is asserting a form of projectivism about value) (Hume, 1975, p. 294):

> The one [reason] discovers objects as they really stand in nature, without addition or diminution: the other [taste] has a productive faculty, and gilding or staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation.[10]

Wittgenstein seems to be saying that what we take to be a necessary truth about the world is really a projection onto the world of an aspect of the rules of our language.[11]

What we have here is something like the classic Humean contrast between realism about the world and anti-realism about modality. In this case, the claim is not that "Upon the whole, necessity is something, that exists in the mind, not in objects," but, in effect, that, upon the whole, necessity is something that exists in linguistic rules, not in objects. And this is not without consequences of its own, of course. What it means is that when someone seems to say something which we might think of as the negation of a necessary truth, it is not that they have said something which is necessarily false, but that they have broken the

rules, and so have really failed to say anything at all: they have just been misusing the words, producing something which is ungrammatical, or nonsense.

What is notable about this view of Wittgenstein's is that there is no detour, as there was in Dummett's case, by way of thought. The source of what seems to us to be necessity in the world is, according to Wittgenstein, nothing but the rules of a particular language; and the key thought here is that these are in a fundamental way arbitrary and historical (Wittgenstein, 2009a, 372). The relevant kind of arbitrariness is just that the rules of a language are not dictated by the way the world is (Wittgenstein, 2009b, 366); so the arbitrariness which Wittgenstein finds in linguistic rules is just another aspect of his anti-realism about necessity – for if linguistic rules had not been arbitrary in this sense, they would have reflected a deeper necessity in the world. Once this is appreciated, there seems an obvious reason to find the source of necessity in language, rather than in thought: the rules of particular languages are more naturally taken to be arbitrary than the rules of thought. If we had taken linguistic rules to be a reflection of the structure of thought, it would have been natural to take them to be determined by the way the world is; it would have been natural to take linguistic rules to be an expression of the necessity of that part of the world which is the way we think.

This fundamentally Humean view of necessity leads naturally to a downgrading of the importance of philosophy, and that is indeed characteristic of the later Wittgenstein. It has recently become common to read a similar attitude to philosophy back into his earlier work,[12] but this is hard to square with some central features of that earlier work. The central claim of the *Tractatus* is that the form of language is the same as the form of the world: in other words, that the ways things can be combined in the world are the same as the ways words can be combined in language. The important thing here is that this is a modal claim: the possibilities for things in the world are the same as the possibilities for elements of language. This immediately makes it impossible to claim that one set of possibilities is less real than the other. The broadly Humean view of the later philosophy, which is naturally characterized as the view that what seems to be necessity in the world is nothing but a projection of the grammar of a language, cannot be made intelligible if the necessity of grammar and the necessity of real combination are the same thing.

In the light of this, people are often tempted to understand the *Tractatus* as a thoroughly realist work: the grammar of language is taken to be a mirror of the form of the world, with the form of the world being imagined to be determinate independently of language (see, e.g., Pears, 1987). Myself, I think this is wrong. Wittgenstein says "what the solipsist *means*, is quite correct" (Wittgenstein, 1922, 5.62), and I see no reason to deny that he is here expressing his own view, even if, strictly speaking, "it cannot be *said*, but it shows itself": in effect, Wittgenstein is a kind of transcendental idealist. It is true that the theory of the *Tractatus* means that nothing can strictly be said about the form of the world – about what is objectively necessary – and the result is that there are almost exactly the same restrictions on the proper range of philosophy as Hume himself would have imposed. But the reason is not, I think – as it seems to have been for Hume – that metaphysical pronouncements are the bogus misrepresentation of features of our system of representation as features of the world, but that what metaphysics would say if it could say anything is too deep in the structure of things to be said. And in any case, there is exactly the same restriction on saying anything about the form of *language* as there is on saying anything about the form of the world – as what I have called the central claim of the *Tractatus* clearly requires. We have a view which looks as if it is Humean, when its spirit is the very opposite of Hume's (see

Carnap, 1963, pp. 25–26). For all that, it does look as if at the heart of the work is an anti-realist view of modality: it is just that – if I am right that Wittgenstein at this point ends up with a form of transcendental idealism – anti-realism about modality leads to a more general anti-realism.

Once again, though – despite Dummett's insistence to the contrary (Dummett, 1978e, p. 442) – there is no detour here through consideration of thought. Thoughts are indeed mentioned (they are introduced at Wittgenstein, 1922, 3, before the official introduction of language at Wittgenstein, 1922, 3.1), but they seem to be kinds of sentence in the mind, rather than the content of sentences (which is what they are for Frege).[13] And at every juncture it is the relation between the world, on the one hand, and *language*, rather than thought, on the other, which is important.[14] Why should this be? It cannot be to support an anti-realist view of necessity by allowing grammar to be arbitrary relative to the world – the view which is to be found in the later philosophy – for in the *Tractatus* neither grammar (here just syntax) nor the form of the world is arbitrary. I suspect that there are two key reasons, both derived ultimately from Frege. First, whatever else it is about, the *Tractatus* is about logic, and the formal representation of logic is one of its chief concerns: this immediately gives an important place to language – or at least to the refined symbolism of the fully analyzed language whose sentences will consist just of names of simple objects. The other reason is the Context Principle. In its loosest forms, this principle says little more than that there is no more to the meaning of a word than its contribution to the meaning of sentences of which it can form part.[15] But Wittgenstein adopts the strictest possible interpretation of it: "only in the context of a proposition [sentence] has a name meaning" (Wittgenstein, 1922, 3.3). The core thought here is that there is something basic about sentences – what distinguishes them from lists – which cannot be captured by thinking of sentences as constructed out of independently intelligible words. It is tempting to think that this inevitably puts language at the center of the picture: sentences simply strike us as units. As for thought and the world, on the other hand, even if we acknowledge counterpart kinds of unit – judgments or thoughts, and facts or states of affairs, respectively – it is only as the counterparts of sentences that they are intelligible as units.

4 Carnap and Quine

Even if it is in fact questionable whether the *Tractatus* is a Humean work, it was certainly taken up in a Humean way – most notably by its most famous philosophical reader, Rudolf Carnap. The striking influence of the *Tractatus* on Carnap is evident if we compare his first book (Carnap, 1967a) with an article published in the same year (Carnap, 1967b). The first version of Carnap (1967a) was written between 1922 and 1925, before Carnap had read the *Tractatus*. Carnap (1967b), however, was written in 1927, after he had read the *Tractatus*. There are two striking differences which are relevant here. First, Carnap (1967a) is not centrally about language: although the book embarks on its project by considering the form of scientific *statements*, and advances further in the same way, what it aims to construct is a system of *concepts* – and hence of their correlative objects – on the basis of a set of fundamental *concepts*. And second, although the book does contain (Carnap, 1967a, §106) the characteristic Humean division between the modal and the world-involving, and with it a rejection of metaphysics as Kant conceives of it, in its main treatment of philosophical issues it is surprisingly permissive, merely insisting that they belong to metaphysics and not to science. Carnap (1967b), however, is much more aggressive. Carnap's dismissal of

metaphysics here does not overtly depend on Hume's Fork, but it does depend on a verificationist criterion of 'factual content' for counting any statement meaningful.[16] And the centrality of that meaning criterion also puts language at the heart of things.

Things are both more explicit and more direct in Carnap (1959), first published just four years later. Here we get an almost explicit restatement of Hume's Fork (Carnap, 1959, p. 77):

> (Meaningful) statements can be divided into the following kinds. First there are statements which are true solely by virtue of their form ("tautologies" according to Wittgenstein; they correspond approximately to Kant's "analytic judgments"). They say nothing about reality … Secondly there are the negations of such statements ("*contradictions*"). They are self-contradictory, hence false by virtue of their form. With respect to all other statements the decision about truth or falsehood lies in the protocol sentences. They are therefore (true or false) *empirical statements* and belong to the realm of empirical science.

And this is applied to dismiss any kind of philosophical statement one might imagine: all that is left of philosophy is a certain method, rather than any body of knowledge. What we have here is the result of a Humean view of necessity, reshaped in line with a Humean reading of the *Tractatus*, to make the crucial relation one between language and the world, rather than between the mind and the world.

This general linguistic reorientation of Humeanism continues into Carnap's mature philosophy, though with an interesting twist. In his (1956a), Carnap introduces the idea of *linguistic structures*, which become *linguistic frameworks* in the more thorough treatment of these issues in his (1956b). A linguistic framework is a language (or part of a language) together with a semantic interpretation of it which determines, among other things, which entities are correlated with its terms.[17] Given the idea of a linguistic framework, Carnap is able to distinguish between what he calls *internal* and *external* questions. Internal questions are questions raised within a linguistic framework, and to these a version of Hume's Fork applies. The semantic rules of the framework determine first that some sentences will be *L-true* – true in the framework "in such a way that [their] truth can be established on the basis of the system … alone, without any reference to (extra-linguistic) facts" (Carnap, 1956a, p. 10) – second, that some sentences will be *L-false* – those whose negations are L-true – and third, for the rest, which empirical circumstances would be involved in their being true. The sentences which are L-true within the system are meant to be those which Kant would have counted as analytic; all other truths are synthetic. External questions, by contrast, are questions about whether to adopt some particular linguistic framework. To this class Carnap consigns questions of ontology, in so far as they are both significant and intelligible. External questions are essentially practical questions in Carnap's view (Carnap, 1956a, p. 43): questions about which framework works best for a given purpose. He thinks of the problems involved here as, in a sense, 'engineering problems' (Carnap, 1956a, p. 43). And he does not assume that the same framework will be appropriate for every purpose (1956a, p. 43; 1956b, p. 208). The important point here is that the question whether to accept a particular linguistic framework *replaces* the question whether the entities referred to by the framework's basic terms really exist: it is not the *same* question. The question whether the entities referred to by a framework's basic terms *really* exist is still regarded – just as it was in Carnap (1967b) – as a pseudo-question (Carnap, 1956b, pp. 214–215). There is no issue of truth for a framework, or of the correctness of accepting it, beyond the question whether the framework is practically useful for a given purpose.

Carnap's general view of philosophy forms the background, and to some extent the target, of the work of W. Van Orman Quine. It is a delicate question exactly what kind of challenge to Carnap's picture is provided by Quine's most famously revolutionary essay, "Two dogmas of empiricism" (Quine, 1953). We can pick out two parts of this essay which are particularly significant for our purposes. First, Quine attacks the notion of analyticity, arguing that it can be given no proper explanation and that its application is unclear. Second, Quine proposes an alternative picture, in which analyticity apparently has no place. The simplest way of understanding what is novel about the new picture is this. First, where Carnap insists on a firm distinction between two kinds of question – internal questions, which operate within the semantic rules of a given framework, and external questions, which concern the choice of framework – Quine makes no such distinction: we have, on the one hand, a single "field of force" which contains "[t]he totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic" (Quine, 1953, p. 42), and on the other hand, experience, which may or may not be recalcitrant; any part of the former may be adjusted to deal with a failure of fit with the latter. Second, where Carnap insists on a firm distinction (within each framework) between those truths which are analytic and those which are not, Quine characterizes analytic truths as those which "hold come what may," and then insists that there is no absolute distinction between those statements which can be held true come what may and those which cannot: "[a]ny statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system"; and "[c]onversely, by the same token, no statement is immune to revision" (Quine, 1953, p. 43).

These ways in which Quine differs from Carnap seem to complicate the relationship between language and philosophy – and indeed between philosophy and other disciplines. Whereas for Carnap there is a relatively clear and separable task for philosophy in the working out of the semantics of linguistic frameworks, and in the higher-order questions of what is involved in that working out, on Quine's account there seems to be no way of separating those questions from other, more obviously scientific, questions, and correspondingly no clear and separable role for philosophy: it is just that the philosopher works in the part of the 'field of force' constituted by the totality of our beliefs which is relatively close to its more theoretical 'core,' at some distance away from the 'periphery' where experience is actually encountered; the philosopher seems to become a kind of theoretical scientist. The abandonment of Hume's Fork in the rejection of the distinction between analytic and synthetic truths seems to bring with it an abandonment of the idea that there is a special relationship between philosophy and language.

That, at least, is how things seem, on the face of it. In fact, however, although there is indeed a radical divergence, it is not clear that it has quite the form which Quine suggests. First, Quine himself seems clearly to endorse a form of linguistic Humeanism in his (only slightly later) treatment of necessity. In (Quine, 1966) Quine attempts to demonstrate that the idea of *de re* modality – of necessities and possibilities arising from the nature of things – is incoherent. But he does not reject necessity altogether, and he clearly identifies the idea of *de re* modality with a pre-Humean – more specifically, Aristotelian – conception of necessity. He thinks we need some notion of necessity to characterize implication, and we need the notion of implication to characterize the notion of validity which we need for proof theory. Much of the technical part of (Quine, 1966) is devoted to the preliminary work of providing the semantics for a notion of necessity which will do this work without bringing in a pre-Humean idea of modality. The idea is to characterize a predicate of

necessity which will apply to whole sentences, without introducing any modality which applies to objects in virtue of their natures. Quine puts the point like this:

> Necessity as a semantical predicate reflects a non-Aristotelian view of necessity: necessity resides in the way in which we say things, and not in the things we talk about. (Quine, 1966, p. 176)

This looks like an echo of Hume (1978, p. 165); it seems inevitably to reintroduce Hume's Fork, and with it some (possibly restricted) notion of analyticity.

A little thought suggests that there is nothing here which is fundamentally incompatible with the holistic 'field of force' picture with which Quine aims to replace Carnap's conception of questions which are internal and external to linguistic frameworks – at least if we correct what looks like a mistake in Quine's presentation. I noted before that Quine characterizes analytic truths as those which "hold come what may," and that he goes on to reject the analytic–synthetic distinction by claiming that "[a]ny statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system." There is an obvious slide here, even if it may seem a more natural slip for an anti-realist about modality to make. The things which 'hold come what may' are the analytic truths themselves (if there are any): to say that they hold come what may is to say that they are true no matter what – in other words, one would think, that they are necessarily true. This is a remark about statements and their status – not a remark about us. But when we say that "any statement can be held true come what may," it is we ourselves who do the holding: the claim Quine is making here is that we can rationally accept or reject any statement at all, provided we make suitable compensations elsewhere. This is a remark about us, and says nothing about the modal status of any statement: in particular, it does nothing to suggest that there are no truths which hold come what may. What this means is that Quine's holistic vision of the totality of our beliefs forming a single 'field of force' is in fact compatible with acknowledging that some truths are necessary and some are not. And, as we have seen, it is at least Quine's only slightly later view that we have to accept this if we are to make sense of implication and validity.

If we incorporate this point, what difference does it make to Quine's view, and in particular to the difference there appeared to be between his view and Carnap's? First, among the beliefs which form Quine's 'field of force,' there may be some beliefs that certain statements are necessary. We have seen that Quine seems later to be committed to allowing this, in order to make sense of implication. And there seems good reason to think that he will need there to be some such beliefs, if he is to make sense of there being any confrontation between the totality of our beliefs and experience. For if our beliefs have no implications, it is hard to see how any experience could be recalcitrant, and so force revision, or how any experience could confirm our beliefs, and so encourage us to leave things unrevised.

This has no tendency to undermine the holism of Quine's general picture, or to suggest that these beliefs about what is necessary will be held true come what may: we can continue to hold that "no statement is immune to revision" – and this will include statements to the effect that this or that other statement is necessary. This suggests that how wide the class of necessary truths is is something which cannot rationally be determined once and for all, in advance of general scientific enquiry: in particular, we cannot rationally restrict it to what we might think of as narrowly logical truths, or to any particular wider class of analytic truths.

What does change, however, is the place and role of philosophy. If the totality of our beliefs "face the tribunal of sense experience not individually but as a corporate body" (Quine, 1953, p. 41), and "no statement is immune to revision," it looks as if the decision to revise a belief about the necessity of some statement – in effect, for Quine, a belief about what implies what – is one that could in principle be made by a scientist, rather than a philosopher. If that is right, the central moral which is to be drawn from the revolutionary new picture of the essay is not anything in particular about analyticity, but a form of the general view with which Quine was later to be associated – what is known as his *naturalism*. The key move here is the "abandonment of the goal of first philosophy" (Quine, 1981, p. 72); he ends up with a view that sees philosophy "not as an *a priori* propaedeutic or groundwork for science, but as continuous with science" (Quine, 1969, p. 126). These remarks seem if anything to understate the change Quine envisages. It is not just that there is no *first* philosophy, on Quine's picture – rather, it seems that there is no distinctive role for philosophy, as we ordinarily envisage it, at all: there is nothing more than very theoretical science.

For all that, in his actual practice Quine seems to have continued very much as Carnap might have (making allowances for other differences between them). His work continues to be focused centrally on the analysis of language and on semantic proposals for linguistic constructions of various kinds. It is no accident that the great monograph of his mature years is called *Word and Object* (Quine, 1960). And in this language-centered work, there is little sense that the semantic claims which are made take into account the latest developments of empirical science.

5 Ordinary Language Philosophy

We have seen that the special place given to the philosophy of language in the logical positivist tradition arises naturally from that tradition's broadly Humean anti-realism about modality. But what are we to make of the anti-theoretical, anti-systematic movement which grew up a little later, known rather loosely for practicing 'ordinary language philosophy'? This movement might seem to place language at the center of philosophy even more obviously than Carnap and Quine, for example, do, but it characteristically avoids the grand methodological visions and manifesto declarations of a Carnap or a Quine; we will not easily find any clear commitment to Hume's Fork here. I will argue that the ordinary language tradition had its *origins*, at least, in anti-realism about modality, and continued throughout its history to take an attitude to philosophy in general, and metaphysics in particular, which is hard to justify without that anti-realism – even if it is characteristic of the philosophers in this tradition that they did not generally attempt to justify it.

The ordinary language tradition is often associated with the work of the later Wittgenstein,[18] and it is in that association that a form of it continues to this day. But it seems more plausible to suggest that its origins lie rather in Wittgenstein's *earlier* philosophy. If anything deserves to be counted as the founding text of ordinary language philosophy, it is Gilbert Ryle's "Systematically misleading expressions" (Ryle, 1932). What Ryle wants to argue for is this claim (Ryle, 1932, pp. 142–143):

> There are many expressions which occur in non-philosophical discourse which … are … couched in grammatical or syntactical forms which are in a demonstrable way *improper* to the states of affairs which they record.

This gives philosophy a task, which Ryle himself is inclined to believe is "the sole and whole function of philosophy" (Ryle, 1932, p. 170): to reformulate the misleading expressions to ensure that "the syntactical form is proper to the facts recorded" (Ryle, 1932, pp. 142–143).

Although Ryle himself disagrees with one important feature of (Wittgenstein, 1922), it seems clear that in identifying "systematically misleading expressions" and reformulating them to avoid what is misleading about them, he is doing just what Wittgenstein seems to be recommending in that work. Ryle identifies the point of disagreement as the view that "what makes an expression formally proper to a fact is some real and non-conventional one–one picturing relation between the composition of the expression and that of the fact" (Ryle, 1932, p. 167), but some notion of "propriety of grammatical to logical forms" there apparently has to be: it is just that at this point Ryle takes it to be "more nearly conventional than natural" (Ryle, 1932, p. 168).

The role given to philosophy by Ryle seems precisely the one assigned to it by Wittgenstein (Wittgenstein, 1922, 4.112): "The object of philosophy is the logical clarification of thoughts." And what this means is made clear by an earlier remark, approving of Russell (Wittgenstein, 1922, 4.0031):

> All philosophy is "Critique of language" (but not at all in Mauthner's sense). Russell's merit is to have shown that the apparent logical form of the proposition need not be its real form.[19]

In (Ryle, 1932), Ryle offers no reason for his view that this clarification of syntax is "the sole and whole function of philosophy." But there is a hint in a later paper – also largely concerned with philosophical methodology – of what lies in the background (Ryle, 1937). Considering the kind of principles which define a school of philosophy, he asks how they are to be established, and it is clear from the reference to Hume and Kant in his answer (Ryle, 1937, p. 325) that he takes Hume's Fork to provide a significant challenge to any account; and as we saw, the problem there derives from anti-realism about modality.

I said that (Ryle, 1932) has a good claim to be the founding text of ordinary language philosophy – the movement in philosophy in which language most obviously takes the center of the stage. There is one way in which it is uncharacteristic of the movement as a whole: at the time he wrote it Ryle says that he was still under the influence of the doctrine "according to which there were a certain number of logical forms which one could somehow dig up by scratching away at the earth which covered them" (Rorty, 1967, p. 305). This doctrine he later abandoned. But in all other respects, it seems to me, the work is broadly representative of the whole movement. We might note the following key points in particular:

(i) It holds that it is at least a central task of philosophy to remove confusions which arise from misunderstanding ordinary language;
(ii) It takes philosophers to be subject to these confusions, while ordinary people are not;
(iii) It is generally pessimistic about the possibility of substantial or systematic philosophy (metaphysics, for example), without offering any principled justification for this pessimism.

These features undergo a slight transformation in the work of J. L. Austin, the other obvious pillar of the ordinary language movement.[20] He is concerned to remove confusions

which arise from misunderstanding ordinary language, but his more central concern is to look attentively, with an almost literary eye, at distinctions which are made in ordinary language, the careful observing of which he thinks will prevent us falling into the kinds of confusion to which philosophers are prone.[21] This seems to me to be broadly continuous with Ryle's concern in (Ryle, 1932), even if its focus tends to be on particular words, rather than on constructions.[22]

Feature (iii) seems to me to be particularly characteristic of the ordinary language movement. The lack of principled justification for the pessimism about substantial or systematic philosophy might seem to undermine my suggestion that the linguistic turn is motivated by a roughly Humean view about necessity, given a linguistic cast by Wittgenstein's early work. But the fact that Ryle's approach has its roots in (Wittgenstein, 1922), and that he was clearly sensitive to Hume's challenge to metaphysics, provides at least genealogical support for my suggestion:[23] the movement's character is still broadly Humean, even if its members were not all Humean in their conscious allegiances. (And in fact I suspect that some form of Humeanism was at least the closet orthodoxy of the age.[24])

The relation of the later Wittgenstein to this school seems to me complicated, despite the fact that he is often associated with it. The later Wittgenstein engages with issues in metaphysics more explicitly than those most obviously within the ordinary language movement – particularly in connection with mathematics. And I have already argued that he is naturally read as a Humean about necessity. But he can be made to seem closer to ordinary language philosophy by concentrating in particular – and often out of context – on those parts of his later texts which seem to exemplify features (i), (ii), and (iii). (See, for example, Wittgenstein, 2009a, 116, 124, 128, 133.)

In a similar way, the early Wittgenstein can be brought broadly within the ordinary language movement by picking out those parts of (Wittgenstein, 1922) which I earlier suggested as an inspiration for Ryle – while ignoring their rationale in a particular metaphysics of modality. (The recent 'new' or 'resolute' reading of (Wittgenstein, 1922) seems to me to do exactly this. See, e.g., Diamond, 1991, and Conant, 2000.) In the case of the early Wittgenstein, this seems to me to lead to a misinterpretation;[25] in the case of the later work, to an oversimplification, at the very least.

6 The Turn Back

I have suggested that the linguistic turn, in all its various manifestations, depends on a general anti-realism about modality, which takes a linguistic form as a result of the early Wittgenstein's concern with formal logic and his commitment to the importance of the Context Principle. Philosophy has moved on from the linguistic turn in its classical forms, however – at least in the English-speaking world: metaphysics is now pursued unapologetically, and central debates in metaphysics have a place in contemporary philosophy not unlike the place occupied by the various manifestations of the linguistic turn in their heyday. If I am right about the basis of the linguistic turn, we should expect the turn back to depend on a rejection of anti-realism about modality.

And so indeed it seems. It is natural to trace the beginning of the gradual rejection of the linguistic turn to the work of Ruth Barcan Marcus, who set about patiently unpicking Quine's arguments against *de re* necessity – his arguments for his version of Hume's Fork.[26] Marcus's work was taken up, developed, transformed, and made readily accessible even to

philosophers with no great technical expertise, by a younger ally in her exchanges with Quine, Saul Kripke.[27] Developing a larger philosophy from Marcus's important thought that names are just tags while descriptions are not, pressing further Marcus's sensitivity to the possibility that different contexts might impose different constraints on substitution, and building on his own earlier work on the semantics of modal logic, Kripke challenged the assumptions on which anti-realism about modality depends, and made the idea of metaphysics respectable again. He insisted on a crucial difference between the notions of the *a priori* and the necessary. The former deals with an issue about how things are known: it is, as Kripke says, "a concept of epistemology" (Kripke, 1980, p. 34). The concept of necessity, on the other hand, "is not a notion of epistemology, but of metaphysics, in some (I hope) non-pejorative sense" (Kripke, 1980, pp. 35–36). This apparently simple point transformed philosophy. There may have been an argument back in the history of philosophy for assimilating the *a priori* and the necessary – perhaps the combination of Hume's empiricism with his theory of perception provides one – but from now on some argument had to be provided: the two concepts could not just be assumed to coincide. As for the concept of analyticity – once the notions of the *a priori* and the necessary have been distinguished, it no longer seems to have such central importance: Kripke simply stipulates that what is analytic is both necessary and *a priori* (Kripke, 1980, p. 39).

Kripke also provides a direct, informal attack on the Humean approach to modality, by showing, with a range of simple examples, how natural and unproblematic in everyday life *de re* modality is.[28] He even turns the ordinary language movement's suspicion of philosophy back against it (Kripke, 1980, p. 41):

> Suppose that someone said, pointing to Nixon, 'That's the guy who might have lost.' Someone else says 'Oh no, if you describe him as "Nixon," then he might have lost; but, of course, describing him as the winner, then it is not true that he might have lost.' Now which one is being the philosopher, here, the unintuitive man?

If Kripke made metaphysics respectable for the majority of philosophers, whether or not their interests lay in formal logic, it is his contemporary, David Lewis, who provides the clearest model of what a modern metaphysics might be like. Lewis's work ranges across many fields – wherever he found a philosophical puzzle, he seems to have offered a novel solution to it – but at its core is his approach to modality, and the distinctive way in which he defends his modal metaphysics. The key to Lewis's conception of modality is his understanding of possible worlds. Possible worlds have a significant history in analytic philosophy: the notion can be traced back to Carnap's notion of 'state descriptions,' which he uses to define the notion of L-truth (Carnap, 1956a, p. 9), and beyond him to Wittgenstein's notion of a 'possible state of affairs' (Wittgenstein, 1922, 4.462).[29] A real transformation occurred, however, in the late 1950s and early 1960s: to simplify hugely, the equivalences between "*s* is necessarily true" and "*s* is true in every possible world," on the one hand, and "*s* is possibly true" and "*s* is true in some possible world," on the other, were exploited to enable the familiar semantics for ordinary predicate logic to be used to generate semantics for modal logic (see in particular Kripke, 1959; 1963a; 1963b).

Lewis's view of modality can be characterized by means of four key commitments (though these are not always clearly separated). First, he is a *realist* about modality: that is, he is committed to the view that claims about what is necessary and what is possible can be strictly true and strictly false, and their truth and falsity is not dependent on facts about

thought or language.[30] Second, he is what may be termed a *literalist* about semantics: he seems to accept the view that it is in principle possible for a semantic theory for a language (or for some part of a language) to be literally true, and if that semantic theory implies a given claim, then that claim is literally true. For our case, if the true semantic theory for modal language (as well, perhaps, as propositional-attitude constructions and talk of properties) implies that there are possible worlds, then there literally are such things as possible worlds. Since Lewis holds that the true semantic theory does indeed involve the claim that there are possible worlds, and he is also a realist about modality, he is committed to the view that possible worlds really exist, independently of facts about thought or language. Third, he is a *reductionist* about modality: that is, he thinks modality – necessity and possibility – needs to be defined in more fundamental terms (Lewis, 1973, p. 85). In particular, he thinks that facts about necessity and possibility are reducible to facts about possible worlds: the facts about possible worlds are fundamental, and can be used to explain necessity and possibility. And fourth, Lewis is a *concretist* about possible worlds: that is, he thinks possible worlds are themselves concrete individuals, spatio-temporal entities "something like remote planets" (Lewis, 1986, p. 2).

These last two commitments are the most contentious features of Lewis's approach to modality. Lewis initially describes possible worlds as "ways things could have been" (Lewis, 1973, p. 84), but it is quite unclear that ways things could have been are concrete individuals (see Stalnaker, 1976). And if we try to reduce necessity and possibility to facts about possible worlds, there is at least some work to be done to explain how the nature of a *possible* world is independently intelligible – intelligible, that is, independently of necessity and possibility.

On the other hand, the first two of the four commitments of Lewis's view which I have identified seem to me to be central to the mainstream of metaphysics after the linguistic turn. I will deal briefly here with the first one – realism about modality – before turning to the second – literalism about semantics – in the next section. I have suggested that it was anti-realism about modality which lay at the root of the linguistic turn which dominated English-speaking philosophy in the middle third of the twentieth century, and which was broadly hostile to metaphysics. I think it is clear that it was the questioning of that anti-realism which opened the way back from the linguistic turn. Not every modern metaphysician need be a realist about modality, even if many indeed are: they might be anti-realist, or they might not be concerned with modality in particular. But the challenge to the anti-realist orthodoxy has meant that people can be metaphysicians again, in ways that reach back to work which was done before the linguistic turn. I began with the strikingly contrasting views of Dummett and Williamson on how philosophy ought to be done. It is no accident that Williamson, after a series of negative arguments against the general conception of philosophy which Dummett might be taken to represent, begins the positive task of sketching an alternative view by presenting the outline of an epistemology of modality (Williamson, 2007, ch. 5). I have suggested that the source of the language-oriented hostility to metaphysics which characterizes the linguistic turn lies in a form of Humeanism about necessity. That Humean view in turn depends on a certain epistemology: empiricism as a general theory of how we can have knowledge of what is independent of us, combined with a particular theory of perception. Overturning the Humean view must involve producing an alternative epistemology of modality, which is exactly what Williamson does.

7 The Larger Picture

Despite the fact that Lewis represents a rejection of the linguistic turn, language is still of fundamental importance to his philosophy. This is because of the semantic literalism which he shares with the mainstream of contemporary metaphysics.[31] I want now to raise some questions about this semantic literalism. The position which contemporary metaphysicians commonly end up with depends on the thought that the best semantic theory for our languages and our logical systems reveals the ultimate nature of reality. This does not in itself mean that semantics should or can be pursued independently of other considerations in order then to reveal the nature of reality, without further adaptation. There need be no bar, in principle, to other, non-linguistic considerations being brought to bear on our choice of semantic theory. Nor, therefore, does semantic literalism *have* to take semantics, or the philosophy of language more generally, to be prior to metaphysics – even if semantic literalists may often be tempted to do so.[32] Of course, since semantics is naturally understood as connecting language with the world, we may anyway think that semantics and metaphysics need to be done together; but the semantic literalist need not then suppose that there is at least something about language which we can get at independently of metaphysics, which decisively determines the truth in metaphysics. Semantic literalism does not, therefore, have to take syntax, for example, to be both prior to and determinative of ontology.[33] Semantic literalism is just the view that we should end up with a semantic theory which connects language and logic, on the one hand, with the fundamental nature of reality, on the other, and that, if we get things right, what such a theory is committed to reveals the literal truth about reality.[34] What this means is that the semantic literalist can be understood as accepting a form of the key assumption of the argument we derived from Dummett at the end of §2:

(4) Metaphysics (or the nearest we can get to it) is to be pursued through the philosophy of language.

It is true that (4) is naturally read as suggesting that we can do the philosophy of language in some way in advance of doing metaphysics, and then derive the metaphysics from it; but it can also be read more neutrally, as suggesting that metaphysics and the philosophy of language must be worked out together.

Semantic literalism applies at different levels, from the more specific to the very general. At the more specific level, we might find, for example, the semantics of claims of the form '*a* knows how to φ' being taken to reveal the true nature of practical knowledge (see, e.g., Stanley and Williamson, 2001, and Stanley, 2011). But it is at the more general level that the issues are most striking. We have seen the suggestion that the best semantic theory for modality shows that there really are possible worlds; but there are more familiar, more apparently venerable examples. The ancient idea that reality contains a fundamental distinction between substance and attribute[35] – or, in its more modern guise, between object and property or relation – seems to imply that the best semantic theory correlates different kinds of entity with singular terms, on the one hand, and predicates, on the other. And another ancient thought, that we should ideally 'carve reality at its joints,'[36] requires reality in itself to have joints; and that in turn seems to need reality itself to be already divided into the classes demarcated by the most fundamental predicates.

There are a number of worries about semantic literalism – though the issues are too large to pursue in any detail here. One reason for worrying about semantic literalism is that

semantic theories are characteristically answerable to considerations which seem not to have much to do with the description of reality. One significant advantage of possible worlds semantics for modality, for example, is that it enables us to use the relatively familiar resources of the semantics of classical predicate logic to derive metalogical results for modal logic. And again, one of the attractions of classical logic itself is the smoothness of the derivation of metalogical results for it. It is not obvious why we should think that these benefits have much to do with the description of the ultimate nature of reality – unless we are either anti-realists, thinking that reality itself is nothing more than a projection of our favored theories, or quite optimistic about the way things are shaped independently of us.

It may be replied to this that similar considerations – of simplicity, for example, or smoothness of application – play an important role in our choice of scientific theories. One response to that comparison might then be to suggest that scientific theories are not themselves concerned just with the description of reality (see, e.g., van Fraasen, 1980). But in fact, there seems to be an important asymmetry between scientific and semantic theories. We expect a scientific theory of some phenomenon to include an account of the relation between the fundamental entities and features posited by the theory, on the one hand, and the phenomenon itself, on the other. But no such thing is generally provided with any seriousness in semantic theories.

At base, semantic theories characteristically *assign* entities or classes of entities to expressions in the language or system in question, and the behavior of the expressions on that assignment is compared with the way the expressions antecedently seem to behave. Deviations between the behavior of the expressions on the relevant assignment and the behavior they antecedently seem to have can be explained in a number of ways. In an ideal case, we might hope to explain all such deviations *away*. What this would leave us with is something like this claim: these expressions behave as they *would* if these assignments *had* been made – it is *as if* these assignments had been made.

Semantic literalism demands something more than this, however: it wants to be able to claim that these assignments have *in fact* been made – the expressions behave like this because these entities or classes of entities *have been* assigned to them. The key move will be a form of inference to the best explanation: the reason why it is *as if* these assignments have been made is because these assignments have *in fact* been made. The difficulty here, however, is that there is seldom much of a story about how the relevant assignments were made. Sometimes there might seem to be a merely local difficulty: a particular semantics-cum-metaphysics might leave us with a puzzling epistemology. (Some of the natural disquiet about Lewis's particular form of realism about possible worlds may perhaps be traced to this.) But there may be a difficulty of principle. If we think that the assignments made by a semantic theory are the same as the assignments which have in fact been made, in some way, in the determination of the meaning of the expressions of a language, it looks as if we need to think of reality as having something which is somehow like the structure which the semantic theory takes the language to have. One difficulty with this is that it is not clear that we can make any significant sense of reality, as it is independently of any relation to language, having structure of the relevant kind. If this kind of difficulty cannot be dealt with, there is a problem with the inference to the best explanation which semantic literalism needs to appeal to: the problem is that the supposed best explanation is no explanation at all. The issue here is exactly analogous to one which arises for a certain form of the argument from design for the existence of God. The relevant form of the argument from design claims that the best explanation of the world being *as if* it had been designed is that it was *in fact* designed. The problem is that we have no idea at all of what the designing

could have been, or what kind of mechanism could have been involved in the execution of the design: with these gaps in the story, it cannot be that the existence of God provides the best explanation of the appearances, because it provides no real explanation at all.

I do not claim that this is a decisive objection to semantic literalism: merely that there is a problem which needs to be addressed. And it is not as if semantic literalism is the only possible approach to semantics. After all, it is already a significant empirical fact about a language that its expressions behave *as if* a certain assignment had been made. It is also a significant empirical fact about a language that its expressions behave in some ways *as if* one kind of assignment had been made, and in some ways *as if* a different kind of assignment had been made. We could here choose to follow Carnap in being tolerant of different semantic theories, each with its own advantages for particular purposes. If we think there are difficulties of principle in accepting semantic literalism, this will not seem like a failure to follow the project of semantics through to its proper conclusion.[37] And we can think there are difficulties with semantic literalism without thinking, as Carnap himself of course did, that metaphysical questions are always pseudo-questions. There may be a proper and significant field of metaphysics, without its being inevitably confined by the framework of semantics.

This clearly makes a difference to the place of the philosophy of language within philosophy as a whole. If we accept semantic literalism, a part of philosophy of language – semantics – remains crucial to metaphysics: even if non-linguistic considerations are taken to constrain what semantic theories we take to be acceptable, metaphysics will always be shaped by the project of providing a semantic theory. If we do not accept semantic literalism, on the other hand, the philosophy of language will not have anything like the same importance, even to metaphysics. (We might think that the philosophy of painting, for example, or the philosophy of music, might be similarly important in giving us a sense of the true nature of reality.)

And of course there is another assumption, which I noted right at the outset, but which has not yet been subjected to any serious examination. It is commonly assumed that metaphysics, in so far as it is intelligible at all, is central and fundamental to philosophy. This was the other key assumption ((1)) of the argument we derived from Dummett. The commonly prevailing general sense of the importance of the philosophy of language for philosophy as a whole seems to derive, because of this assumption, from its relation to metaphysics – either as the key to metaphysics, or as what has to be done in place of metaphysics. But in a more inclusive age, this assumption might also be questioned, and branches of philosophy for which metaphysics seems to have only marginal importance might be given equal respect. In that case too, we might take language to be of constant interest – but perhaps of no more importance than other, quite different forms of representation.

Notes

1 The term appears to originate in Bergmann (1967), though it was made famous by Rorty (1967).

2 For those among whom Williamson is naturally placed: Williamson himself mentions Saul Kripke, David Lewis, Kit Fine, Peter van Inwagen, and David Armstrong (Williamson, 2007, p. 19).

3 This suggestion was made to me in conversation by Barry Smith.

4 It is beyond the scope of a footnote to make this claim out in detail: for sketchy support for it I note the apparently verificationist conception of character found early on in Dummett (1978c, pp. 15–16), and the Humean approach to causation in Dummett (1978b). It is also clear that Dummett took the later work of Wittgenstein very seriously – see, e.g., Dummett (1978d, p. 436): the significance of that will appear shortly.

5 The suggestion that this is the ultimate source of Dummett's outlook was made in conversation by Michael Rosen. I here adopt a fairly traditional reading of Hume.

6 The reason for the caveat is the general anti-realism which seems to follow from other things Hume says (Hume, 1978, pp. 66–68): that is, Hume is at least more realist about the world than he is about, say, value (or necessity). But I will not complicate the issue in the main text.

7 For the classic statement see Hume (1975, pp. 163–165), though there is an earlier version in Hume (1978, pp. 413–414).

8 I understand realism about modality, here and throughout, as realism about necessity and possibility, rather than as the specific form of realism about possible worlds which is associated with David Lewis – though I will mention that specific view below.

9 We might note in passing, against those who think that Wittgenstein only ever wanted to say what everyone would agree with, that this modal claim coincides with his own philosophy, but conflicts with what almost everyone regards as common sense, as Wittgenstein himself was clearly aware (Wittgenstein, 2009a, 246).

10 We see Wittgenstein himself adopting a form of projectivism about clear and strict rules of syntax in Wittgenstein (2009a, 104): "One predicates of the thing what lies in the mode of representation."

11 There is an issue whether the resultant view should be counted as an 'error' theory (like the view of ethics to be found in Mackie, 1977, ch. 1) or not. On an error-theoretic account, claims like 'I can know what someone else is thinking, not what I am thinking' would either assert or presuppose that they concerned some necessity which is in the world independently of language; in that case, if projectivism is true, all such claims would be either false or lacking in truth-value. My own feeling is that Wittgenstein is better understood as thinking that there is nothing wrong with the claims themselves – the error belongs just to philosophers, who simplistically understand these claims on the model of other claims, which are really quite different. Whether this makes Wittgenstein precisely a 'quasi-realist,' in the sense of Blackburn (1984) is another matter. (As Blackburn conceives of the position, 'quasi-realism' is concerned not just to assert a projectivist view without accepting an error theory: it aims to explain and justify the realistic appearance of the language; Blackburn 1984, pp. 171, 180.)

12 This is the so-called 'new' or 'resolute' reading of the *Tractatus*: see, e.g., Diamond (1991) and Conant (2000).

13 To see this point, we need to see that when Wittgenstein first introduces the idea of representing the world for ourselves, he uses the colloquial German "Wir machen uns Bilder der Tatsachen" (1922, 2.1), which might ordinarily be translated "We picture facts to ourselves" (as Pears and McGuinness do; Wittgenstein, 1961, 2.1), but he understands it fully literally, as 'We make to ourselves pictures of facts' (as Ogden has it; Wittgenstein 1922, 2.1). (That he understands it fully literally can be seen from a remark he makes about the parallel passage at Wittgenstein, 1922, 3.001, in a letter to Ogden; Wittgenstein, 1973, p. 24.) And then at Wittgenstein (1922, 3) he says, "The logical picture of the facts is the thought," which makes clear that a thought, like a sentence, is a picture – rather than the content of a picture. Again, at Wittgenstein (1922, 4) he writes, "The thought is the significant proposition [*Satz*: I would generally translate this *sentence* in this text]." And in response to Bertrand Russell's request for illumination about Wittgenstein (1922, 3) – "what are its [the thought's] constituents and components, and what is their relation to those of the pictured Tatsache [fact]?" (Wittgenstein, 1995, p. 121) – Wittgenstein says, "I don't know *what* the constituents of a thought are but I know *that* it must have such constituents which correspond to the words of Language. Again the kind of relation of the constituents of thought and of the pictured fact is irrelevant. It would be a matter of psychology to find out" (Wittgenstein, 1995, p. 125). The fact that this is supposed to be the business of *psychology* makes it clear that thoughts here are mental entities of some kind, rather than the contents of anything mental. Further, in response to Russell's question about Wittgenstein (1922, 4) – "Does a Gedanke [thought] consist of words?" (Wittgenstein, 1995, p. 122) – Wittgenstein says, "No! But of psychical

constituents that have the same sort of relation to reality as words. What those constituents are I don't know" (Wittgenstein, 1995, p. 125).

14 Wittgenstein (1922, 5.62) provides an obvious example.

15 Frege himself gives two slightly different formulations (Frege, 1953, pp. x, 73).

16 Carnap himself attributes the view that "metaphysical sentences are meaningless since they are in principle unverifiable" to Wittgenstein (Carnap, 1967a, p. xi).

17 This is clear from Carnap's discussion in his (1956a, pp. 43–44).

18 See the diagram in Hacker (2013, p. 947), and, with a qualification noted in fn. 23 below, Urmson (1967, p. 297).

19 It is surely no accident that two of Ryle's three classes of "sytematically misleading expressions" are found to be misleading in ways that Russell had earlier anticipated (Russell, 1905). Ryle's language, however, is closer to Wittgenstein's than to Russell's, and Russell did not think philosophy was just 'critique of language.'

20 I am reluctant to include P. F. Strawson here, because he is more nearly a Kantian – more nearly a metaphysician. It is true that the metaphysics is said to be merely "descriptive" (Strawson, 1959, pp. 9–12), but I think this undersells what he actually does.

21 Almost all of Austin's works provide examples of this: Austin (1979) is perhaps the most self-conscious methodologically; and Austin (1962) applies the method most clearly to present a sustained attack on a particular philosophy.

22 Urmson (1967) distinguishes what he calls "Oxford analysis" – by which he seems to mean *Austinian analysis* – more sharply from Rylean or Wittgensteinian analysis than I do.

23 In the sense of Nietzsche (1994).

24 Note, e.g., that in claiming Strawson's "descriptive metaphysics" (Strawson, 1959, pp. 9–12) for the ordinary language movement, Hacker says, "descriptive metaphysics is just more analytic description of the structure of our conceptual scheme, not synthetic description of the structure of the world" (Hacker, 2013, p. 941).

25 This is argued more thoroughly in Morris and Dodd (2009).

26 Above all in Barcan Marcus (1993b).

27 Kripke's deep understanding of the issues is apparent already in the discussion of Barcan Marcus (1993b), which appears in Barcan Marcus (1993a, pp. 24–35).

28 The general point about the everydayness of essentialist talk is made in Barcan Marcus (1993c, p. 55).

29 And, of course, beyond them to Leibniz.

30 Note that realism about modality is implied by, but does not imply Lewis's well-known "realism about possible worlds" – the large view which is announced in Lewis (1973, ch. 4) and defended in Lewis (1986). This latter thesis also involves what I call *concretism* about possible worlds (see below) – and perhaps what I call *reductionism* about modality in addition.

31 We should mention here the work of Donald Davidson, who is not mentioned in the main text, but who was an influential figure at the beginning of the last quarter of the twentieth century, and for whom language was of central importance (see almost everything in Davidson, 1984). I take Davidson to be a kind of bridge figure: carrying on some of Quine's commitments (to extensionalism, for example, which itself looks as if it depends on a form of Humeanism); but also with a cultivated urbanity which made him accessible to the ordinary language tradition; and throughout adopting a form of semantic literalism.

32 So semantic literalism is not committed to what Heather Dyke calls the "representational fallacy," which she describes as "a general strategy of reading metaphysics off language" (Dyke, 2007, p. 7).

33 So semantic literalism is not committed to what Crispin Wright calls "the thesis of the priority of syntactic over ontological categories," according to which "the question whether a particular expression is a candidate to refer to an object is entirely a matter of the sort of syntactic role which it plays in whole sentences" (Wright, 1983, p. 51).

34 See Williamson (2007, p. 142). This approach evidently lies behind Williamson (2013).
35 This evidently dates back to Aristotle (1949). A recent exponent is Lowe (2006).
36 For the origin of the idea, see Plato (1901, 265e). For a contemporary use of it, see Sider (2011).
37 This is the general tenor of the argument of Keefe (2000, pp. 49–61).

References

Aristotle. 1949. "Categoriae." In *Aristotelis categoriae et liber de interpretatione*, edited by L. Mino-Paluello. Oxford: Oxford University Press.

Austin, J. L. 1962. *Sense and Sensibilia*, edited by G. J. Warnock. Oxford: Oxford University Press.

Austin, J. L. 1979 (1956). "A plea for excuses." In *Philosophical Papers*, 3rd edn, edited by J. O. Urmson and G. J. Warnock. Oxford: Oxford University Press.

Barcan Marcus, R. 1993a. *Modalities*. Oxford: Oxford University Press.

Barcan Marcus, R. 1993b (1961). "Modalities and intensional languages." In Barcan Marcus, 1993a, pp. 3–23.

Barcan Marcus, R. 1993c (1971). "Essential attribution." In Barcan Marcus, 1993a, pp. 53–70.

Bergmann, G. 1967. "Logical positivism, language, and the reconstruction of metaphysics." In *The Linguistic Turn: Recent Essays in Philosophical Method*, edited by R. Rorty, pp. 63–71. Chicago: University of Chicago Press. Reprinted (in a truncated form) from *Rivista Critica di Storia della Filosophia*, 8(1953): 453–481.

Blackburn, S. 1984. *Spreading the Word*. Oxford: Oxford University Press.

Carnap, R. 1956a. *Meaning and Necessity*, 2nd edn. Chicago: University of Chicago Press.

Carnap, R. 1956b: "Empiricism, semantics, and ontology." In Carnap, 1956a, pp. 205–221.

Carnap, R. 1959 (1932). "The elimination of metaphysics through logical analysis of language," translated by A. Pap. In *Logical Positivism*, edited by A. J. Ayer, pp. 60–81. New York: Free Press.

Carnap, R. 1963: "Intellectual autobiography." In *The Philosophy of Rudolf Carnap*, edited by P. Schilpp, pp. 3–84. La Salle, IL: Open Court.

Carnap, R. 1967a (1928). *The Logical Structure of the World*, translated by R. George. London: Routledge and Kegan Paul.

Carnap, R. 1967b (1928). "Pseudoproblems in philosophy." In *The Logical Structure of the World*, translated by R. George, pp. 301–343. London: Routledge and Kegan Paul.

Conant, J. 2000. "Elucidation and nonsense in Frege and early Wittgenstein." In *The New Wittgenstein*, edited by A. Crary and R. Read, pp. 174–217. London: Routledge.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Diamond, C. 1991. "Throwing away the ladder: how to read the *Tractatus*." In *The Realistic Spirit*, pp. 179–204. Cambridge, MA: MIT Press.

Dummett, M. 1978a. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. 1978b (1954). "Can an effect precede its cause?" In Dummett, 1978a, pp. 319–332.

Dummett, M. 1978c (1959). "Truth." In Dummett, 1978a, pp. 1–24.

Dummett, M. 1978d (1960). "Oxford Philosophy." In Dummett, 1978a, pp. 431–436.

Dummett, M. 1978e (1975). "Can analytical philosophy be systematic, and ought it to be?" In Dummett, 1978a, pp. 437–458.

Dummett, M. 1993. *Origins of Analytical Philosophy*. London: Duckworth.

Dyke, H. 2007. *Metaphysics and the Representational Fallacy*. New York and Abingdon: Routledge.

Frege, G. 1953. *The Foundations of Arithmetic*, 2nd edn, translated by J. L. Austin. Oxford: Blackwell.

Frege, G. 1977. "Thoughts." In *Logical Investigations*, translated by P. T. Greach and R. T. Stoothoff, pp. 1–20. Oxford: Blackwell.

Hacker, P. M. S. 2013. "The linguistic turn in analytic philosophy." In *The Oxford Handbook of the History of Analytic Philosophy*, edited by M. Beaney, pp. 926–947. Oxford: Oxford University Press.

Hume, D. 1975 (1777). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn, edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Oxford University Press.

Hume, D. 1978 (1739–1740). *A Treatise of Human Nature*, edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Oxford University Press.

Kant, I. 1997 (1781 and 1787). *Critique of Pure Reason*, translated by P. Guyer and A. Wood. Cambridge: Cambridge University Press.

Keefe, R. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Kripke, S. 1959. "A completeness theorem in modal logic." *Journal of Symbolic Logic*, 24(1): 1–14.

Kripke, S. 1963a. "Semantical analysis of modal logic I: normal modal propositional calculi." *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9: 67–96.

Kripke, S. 1963b. "Semantical considerations on modal logic." *Acta Philosophica Fennica*, 16: 83–94.

Kripke, S. 1980 (1972). *Naming and Necessity*. Oxford: Blackwell.

Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lowe, E. J. 2006. *The Four-Category Ontology*. Oxford: Oxford University Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin.

Morris, M., and J. Dodd. 2009. "Mysticism and nonsense in the *Tractatus*." *European Journal of Philosophy*, 17(2): 247–276.

Nietzsche, F. 1994 (1887). *On the Genealogy of Morality*, edited by K. Ansell-Pearson, translated by C. Diethe. Cambridge: Cambridge University Press.

Pears, D. 1987. *The False Prison*, vol. I. Oxford: Oxford University Press.

Plato. 1901. "*Phaedrus*." In *Platonis Opera*, vol. II, edited by J. Burnet. Oxford: Oxford University Press.

Quine, W. V. O. 1953 (1951). "Two dogmas of empiricism." In *From a Logical Point of View*, 20–46. Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1966 (1953). "Three grades of modal involvement." In *The Ways of Paradox and Other Essays*, pp. 158–176. Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1969. "Natural kinds." In *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. O. 1981. *Theories and Things*. Cambridge, MA: Harvard University Press.

Rorty, R. 1967. *The Linguistic Turn: Recent Essays in Philosophical Method*. Chicago: University of Chicago Press.

Russell, B. 1905. "On denoting." *Mind*, 14(56): 479–493.

Ryle, G. 1932. "Systematically misleading expressions." *Proceedings of the Aristotelian Society*, 32: 139–170.

Ryle, G. 1937. "Taking sides in philosophy." *Philosophy*, 12(47): 317–332.

Sider, T. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Stalnaker, R. 1976. "Possible worlds." *Noûs*, 10(1): 65–75.

Stanley, J. 2011. *Know How*. Oxford: Oxford University Press.

Stanley, J., and T. Williamson. 2001. "Knowing how." *The Journal of Philosophy*, 98(8): 411–444.

Strawson, P. F. 1959. *Individuals*. London: Methuen.

Urmson, J. O. 1967. "The history of analysis" (with discussion). In *The Linguistic Turn: Recent Essays in Philosophical Method*, edited by R. Rorty, pp. 294–311. Chicago: University of Chicago Press.

van Fraasen, B. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, T. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*, translated by C. K. Ogden. London: Routledge and Kegan Paul.

Wittgenstein, L. 1961. *Tractatus Logico-Philosophicus*, translated by D. F. Pears and B. F. McGuinness. London: Routledge and Kegan Paul.

Wittgenstein, L. 1973. *Letters to C. K. Ogden, with Comments on the English Translation of the* Tractatus Logico-Philosophicus, edited by G. H. von Wright. Oxford and London: Blackwell, Routledge, and Kegan Paul.

Wittgenstein, L. 1995. *Ludwig Wittgenstein: Cambridge Letters – Correspondence with Russell, Keynes, Moore, Ramsey and Sraffa*, edited by Brian McGuinness and Georg Henrik von Wright. Oxford: Blackwell.

Wittgenstein, L. 2009a (1953). *Philosophical Investigations*, rev. 4th edn, edited by P. M. S. Hacker and J. Schulte, translated by G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. Chichester: Wiley-Blackwell.

Wittgenstein, L. 2009b. "Philosophy of psychology: a fragment." In Wittgenstein, 2009a, pp. 182–243 (1953).

Wright, C. 1983. *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

2

Meaning and Truth-Conditions: From Frege's Grand Design to Davidson's

DAVID WIGGINS

1. However close it may have lain beneath the surface of some earlier speculations about language, the idea that to understand a sentence is to have grasped its truth-condition was first made explicit by Frege, for whom it was simply an unemphasized consequence of his general approach to questions of meaning. In the transition from logical positivism to modern analytical philosophy, the idea came near to being mislaid entirely. It was brought back into a new prominence in the late 1960s by Donald Davidson. Having rediscovered the idea for himself and in his own way, Davidson pressed its claims as a principle in the philosophy of mind and meaning, and as the only proper basis on which to conduct serious semantic investigations.

In advance of considering these and more recent claims about meaning, it will be useful to mark certain moments in the formulation and reformulation of the original insight of the truth-conditional theory. In a historical framework, even the bare skeleton of one furnished here, truth-conditional notions may be expected to transcend our more immediate sources of theory or doctrine concerning them as well as our more ephemeral disputations.

2. What is it for a declarative sentence to mean something, or have a sense? For Frege, to answer such a question was not, as it was later for Carnap or his inheritors, an all-important end in itself. Nor was answering that question a part of a comprehensive effort to arrive at a philosophical account of the relation of language to mind, as it is for Davidson and his inheritors. For Frege, it was a means, a propaedeutic for the understanding of the specific thing whose status and nature centrally concerned him, namely arithmetical judgments. Nevertheless, despite the special character of this original interest, Frege saw the question of the meaning of a declarative sentence as a general question, requiring not so much the introduction of a *calculus ratiocinator* (he said) as the creation of something more resembling a Leibnizian *lingua characteristica*. ("My intention was not to represent an abstract logic in formulas, but to express a content through written signs in a more precise and clear way."[1])

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

What Frege took the answer to his question to require was a general notion of meaning that could be correlative with the general idea of the understanding of a sentence. The conception he formed was of the *Sinn* or *sense* of a sentence that was to be understood thus or so, the sentence itself being seen as something built up by iterable modes of combination from component words, each of which had its own contributory sense. The senses of part and whole were to be such that the latter could be determined from the former (given an account of the modes of grammatical combination involved in the construction of the sentence).

The culmination of Frege's efforts may be found in Volume 1, §32 of the *Grundgesetze der Arithmetik*,[2] where he declares that there is both sense and reference for every sentence of his 'concept-writing' or 'ideography,' his *Begriffsschrift*. The *Begriffsschrift* is the constructed language whose operations are to shadow the workings of natural language and, in matters of difficulty such as the foundations of arithmetic, to regulate or supplant natural language. The reference of a sentence of *Begriffsschrift* is its truth-value, and the sense of the sentence is the thought that the sentence expresses.

But how exactly does a thought attach to a sentence? And what is a thought? Well, which thought it is that a sentence expresses and how the thought attaches to the sentence will depend upon nothing other than this: under what conditions is the sentence to count as true? Or, as Frege describes the matter for the artificial language he has just finished constructing:

> It is determined through our stipulations [for the linguistic expressions and devices comprising the language of *Begriffsschrift*] under what conditions [any sentence of *Begriffsschrift*] stands for the True. The sense of this name [of a truth-value, i.e. the sense of this sentence], that is the thought, is the sense or thought that these conditions are fulfilled…. The names [expressions], whether simple or composite, of which the [sentence or] name of a truth-value is constituted contribute to the expression of a thought, and this contribution [of each constituent] is its sense. If a name [expression] is part of the name of a truth-value [i.e. is part of a sentence], then the sense of the former, the name [expression], is part of the thought expressed by the latter [the sentence].

This statement comes at the end of Frege's detailed explanations of *Begriffsschrift*. But its import is potentially perfectly general, and the stipulations of sense for the expressions of his invented language simulate what it is for the expressions of a natural language to have a given or actual (not merely stipulated) sense. The institution of the *Begriffsschrift* – the project Frege had begun in preparation for his books on the foundations of arithmetic (1884; 1893) and published in part in 1879,[3] but then resumed and substantially corrected in the work of 1893, from which we have just quoted – at once illuminates natural language, albeit only in microcosm, and extends it. It illuminates it by displaying clearly the workings of a distinct language abstracted from natural language, namely the concept-script in which Frege hoped to make newly perspicuous all questions of "inferential sequence." The purposes this serves are akin to the practical and theoretical purposes that the construction of an artificial hand with a specialized function might have for a community of beings whose normal members had natural hands with less specialized functions.

3. Given Frege's concern with "a formula-language for pure [i.e., non-empirical] thought," it is unsurprising that, as he said, he "confined [him]self for the time being to expressing [within it] relations that are independent of the particular characteristics of objects"

(*Begriffsschrift*, 1879, preface). Properties and relations that were not so independent registered in the *Begriffsschrift* only in the form of generality-indicating letters such as Φ or Ψ that prescinded from all particular content.[4] Nevertheless, Frege did envisage successive relaxations of this ordinance, and he spoke of possible extensions of his formula language to embrace the sciences of geometry, motion, mechanics, and so on.

Given the universality and generality of the insights that originate with Frege, what we now have to envisage is a further extension of *Begriffsschrift*, namely the extension which, for purposes rather different from Frege's, will even furnish it with the counterpart of such ordinary sentences as "the sun is behind cloud" (say). In the very long run, the extended *Begriffsschrift* (Bg^+) might itself be modified yet further, to approximate even more closely to the state of some natural language. In the interim, however, in the transition from Frege's to our own purposes, it stands as an illustrative model of something much more complicated.

In an extension such as we are to imagine, a sentence like "the sun is behind cloud" will have a sense if and only if it expresses a thought. For the particular thought that the sun is behind cloud to attach to this English sentence (for it to attach to such a social artefact as this, produced and held fast in its temporal, historic, and social setting, Frege need not forbid us to say) will be for the sentence to be so placed in its total (historical and customary-cum-linguistic) context that it stands [in some situation] for the True just in case [in that situation] the sun is behind cloud. Putting the matter in a way that is not Frege's, one is tempted to say that he who understands the sentence is party to a practice that makes this the circumstance under which the sentence counts as true.

What mystery remains about what a thought is? The thought expressed by a sentence is expressed by it in virtue of ordinary linguistic practices (the practices that we have imagined will be encapsulated in the definitions or elucidations of the empirical terms to be introduced into the extended *Begriffsschrift*), which expose the sentence to reality, and its author to the hazard of being wrong, in one way rather than another way. Once you know that, you know what the thought is that sentence expresses.

4. The truth-conditional thesis, so seen, can be detached from more questionable features of Frege's semantical doctrine, such as the idea that a sentence is a complex sign standing for objects called the True or the False, or is a name of a truth-value. Wittgenstein does detach it (an act of retrieval for which he is too rarely commended) in *Tractatus Logico-Philosophicus* (1921):[5]

> 4.022 A sentence in use (Satz) shows how things stand if it is true. And it says that they do so stand.
>
> 4.024 To understand a sentence in use means to know what is the case if it is true.
>
> 4.062 A sentence in use is true if we use it to say that things stand in a certain way, and they do.

These are striking formulations, more general than Frege's and not radically dependent upon Wittgenstein's picture-theory of meaning. But now it seems we must attend to a problem that neither Frege nor Wittgenstein addressed explicitly. It is the problem (which still excites controversy in connection with Donald Davidson's version of the truth-conditional view of meaning[6]) that not just any true equivalence in the form [*s* is true if and only if *p*] can suffice to show that *s* actually *means* that *p*.

Suppose that the sentence "the sun is behind cloud" is now true. Then all sorts of other things have now (as matters stand) to be the case. It is daytime, the sun has risen, it is not

dark, more people are awake than asleep, and millions of automobiles are emitting smoke into the atmosphere, and so on – all this in addition to the sun's being behind cloud. For these are the accompaniments, in the world as it is, of its being daytime and the sun's having actually risen (to be obnubilated or not obnubilated). It is only to be expected, then, that, where *s* makes such a particular historical statement as it does, in a manner dependent upon some historical context, any of these extra things may in that context be added *salva veritate* to the right-hand side of the biconditional "*s* is true if and only if the sun is behind cloud and…." (It is certain that, without detriment to the truth of *s*, any necessary truth or natural law can be added so.) It is only by virtue of *knowing already* what *s means* that one would pick on the "sun is behind cloud" conjunct, from out of the mass of things that also hold when the "sun is behind cloud" is true, to be the clause to give the proper truth-condition for *s*. It follows that, to put down what a given utterance of a sentence *s* means and impart its meaning to someone, we need to be in a position to signal some 'intended' or 'privileged' or 'designated' condition on which its truth depends. Only where '*s* is true iff *p*' signals on its right-hand side an intended, privileged, or designated condition, can we conclude from this biconditional's obtaining that the utterance of *s* actually means that *p*. Look again at Wittgenstein's 4.024. His "what is the case if it [the sentence] is true" *presupposes* the identification of that intended or designated condition.

5. One way to advance might be to recast Frege's and Wittgenstein's thesis as follows:

> Sentence *s* has as its use to say that *p* – or *s* means that *p* – just if whether *s* is true or not *depends specifically upon* whether or not *p*.

But this is not really the end of the difficulty. For one of the things that the truth of "the sun is behind cloud" (as said at a given particular time and place) depends specifically upon, in one ordinary and standard sense of "depend," might perhaps (at that time and place) be low atmospheric pressure plus the obtaining of other meteorological conditions. None of this, however, is what the sentence actually says. And for the same reason we cannot improve the formulation just given by ruling that the truth of the sentence has to depend *only* upon the designated condition. It cannot depend 'only' on that condition, in the ordinary sense of "depend." For it will have to depend (in that ordinary sense) on everything that the satisfaction of the intended condition itself depends upon.

6. Consider now what Frege could have said in reply to this sort of difficulty, pointing to things already done in *Grundgesetze*. Suppose (as before), that the language of his *Begriffsschrift* has been formally expanded to enable one to say "the sun is behind cloud" and all sorts of similar empirical things. Let us call the imaginary extension (Bg^+). Each new primitive expression ('sun,' 'cloud,' etc.) will have had a reference stipulated for it in accordance with an empiricized extension of Frege's canon for definitions (see *Grundgesetze*, 1893, I, §33). In each case, the sense of the new primitive expression will consist in the fact that its reference is stipulated thus or so.[7] By virtue of this, it will have been contrived that the sense of any complex expression can be determined from its structure and from the referential stipulations governing each constituent expression. But now, in the light of all this, Frege is entitled to insist that, if we stick scrupulously to what actually flows from the full and appointed referential stipulations for all the individual expressions and devices of the extended *Begriffsschrifft* – let us call the set that consists of these stipulations $\Theta(Bg^+)$ – then we shall never arrive at an unwanted

biconditional such as 'the sentence "the sun is behind cloud on 25 June 1993" is true if and only if on 25 June 1993 the sun is behind cloud and the sun has risen and there is low pressure and more people are awake than asleep and …' (or its counterpart in Bg^+). For the stipulations for the extended *Begriffsschrift* furnish no way to derive such a biconditional. The intended condition will be the particular condition that the appointed stipulations deliver. Not only that. In concert, these stipulations, which license nothing about low pressure as part of the truth-condition for *s*, will *spell out* the specific particular dependence that had to be at issue in the restatement of the Frege–Wittgenstein thesis.

No wonder (Frege's philosophical champion may want to say) that we can hear '"the sun is behind cloud" is true if and only if the sun is behind cloud' as more or less equivalent to 'The truth of "the sun is behind cloud" semantically depends upon whether or not the sun is behind cloud.' For we can hear the biconditional '"The sun is behind cloud" is true if and only if the sun is behind cloud' as something delivered to us by whatever plays the part for English that the Fregean stipulations $\Theta(Bg^+)$ will play for the extended *Begriffsschrift*. Where the turnstile "$\vdash$" adjoined to '$\Theta(Bg^+)$' signifies that the stipulations Θ suffice to derive that which follows, what we are saying is, in effect, this:

[*s* means in Bg^+ that *p*] is equivalent to $\vdash_{\Theta(Bg+)}$ [True *s* if and only if *p*].

There is nothing strange or scandalous in the suggestion that we hear the conditional as enclosed or nested in this way within an operator "$\vdash_{\Theta(Bg+)}$" whose presence has to be understood. Countless conditionals we utter are intended by us to be understood as presupposing some norm or tendency that we could roughly identify but do not attempt to describe in the form of an explicit generalization. In so far as some residue of a philosophical problem still persists, the place to which it escapes is the characterization of "$\vdash$" and the general idea of a set of specifically referential specifications that imply this or that equivalence in the form [True *s* if and only if *p*]. The point that is left over, which we shall have to attend to in due course, is that, even though $\Theta(Bg^+)$ would *exemplify* such a set, $\Theta(Bg^+)$ could scarcely stand in for a general characterization of what a referential specification *is*. We need $\vdash_{\Theta(L)}$ for variable L.

7. The residual problem is philosophical – and serious. There is no immediate solution. We will return to the matter in §18 following, at a point decades later in our narrative. Meanwhile, let us consolidate the position now arrived at and pause here to show – if not in Frege's symbolism (which continues to daunt typesetters and readers equally) nor in strict accordance with every particular of Frege's own view of predication[8] – how, more exactly and in more detail, the claim might be made good that Frege can pick out the particular sort of dependence that he needs to secure between the truth of a particular sentence *s* of a language L and the obtaining of some condition that *p*. Let us do so by giving the referential specification of the semantics of a tiny sub-language L(1) of English that might be the counterpart of some small fragment of the extended *Begriffsschrift*.

Suppose the constituent strings of L(1) are simply the following:

(1) The sun is behind cloud
(2) Not (the sun is behind cloud), [which is said aloud as follows: the sun is not behind cloud]

(3) The moon is behind cloud
(4) Not (the moon is behind cloud), [which is said aloud as follows: the moon is not behind cloud],

together with all possible conjunctions of (1), (2), (3), and (4). Then we can determine the sense of an arbitrary string of L(1) by the following provisions:

Terms:	T(1)	"The sun" is a term and stands for the sun.
	T(2)	"The moon" is a term and stands for the moon.
Predicates:	P(1)	"Behind cloud" is a predicate and stands for being behind cloud.
Connectives:	C(1)	"Not" is a unary connective: where A is a string of L, "not" + A is true if and only if A is not true.
	C(2)	"And" is a binary connective: [A + "and" + B] is true if and only if A is true and B is true.
Syncategorematic Expressions:		"Is" is a syncategorematic expression, whose role it is to signal the fundamental mode of combination exemplified in R(1) below.
Rule of Truth:	R(1)	A sentence that is of the form [t + "is" + F], i.e., a sentence consisting of a term t, such as "the sun" or "the moon," followed by the syncategorematic expression, "is," followed by a predicate expression, F, such as "behind cloud," is true if and only if what t stands for has what F stands for[9] [that is to say that the reference of t has the property that F stands for].

Now let us put these rules together and note their effect. Given the sentence ["the moon" + "is" + "behind cloud"] = [The moon is behind cloud], we can agree, by R(1), that the sentence is true if and only if what "the moon" stands for has what "behind cloud" stands for, which last we can show to be true (see T(2) and P(1)) if and only if the moon is behind cloud. That does not make news – no more than news is made when, having multiplied 13 by 25 and got 325, you then divide 325 by 13 and get 25. But it verifies something. Similarly, as Davidson would point out here on Frege's behalf, our semantic derivation helps verify something, namely that, so far as they go, T(1), T(2), P(1), C(1), C(2), and R(1) represent a correct reckoning of the semantic resources of L(1).

What is achieved would have looked more impressive, no doubt, if L(1) had been a fragment of Chinese or Arabic and our referential specification had been done in English. Such a specification is something we can more easily imagine someone's failing to get right. There is no question, however, of a specification of this sort's looking impressive – or its needing to do so – unless it solves neatly and correctly a known grammatical difficulty or casts some light, however indirect, on a real obscurity in the workings of the language under analysis. The specification simply leaves nothing to chance in the idea that, where *s* is an L(1) sentence, *s* means in L(1) that *p* if and only if the biconditional [True *s* if and only if *p*] flows from $\Theta^{L(1)}$. In the context of Frege's own particular purposes in the *Grundgesetze der Arithmetik*, let this serve as a model for the defense of what Frege wanted to say there about

sentence sense. For all he needed to claim at 1.32 of that work was his complete control over the sense of a *Begriffsschrift* sentence. There is no relevant doubt, either theoretical or practical, of that grasp.

8. In *Tractatus* 4.024 Wittgenstein is heir to Frege's idea of sentence sense, and he tries to prescind from the particularities of *Begriffsschrift* in order to make a general claim. Nearby, at 4.002, he points to the need to bring real, live speakers into the picture. Once we take their presence seriously, however, we shall notice for ourselves a new kind of difficulty – the first of several that will come in due course to occupy us.

Consider the Latin sentence *alea jacta est*. Like its standard translation into English, *the die is cast*, the sentence is true if and only if a die [the die] has been thrown. This requires, *inter alia*, that there be a real die and someone who has thrown it. But it is safe to say that what speakers have normally used the Latin or the English sentence to state or to intimate – to say in the full and ordinary sense of 'say' – is nothing of that sort. The normal use of the sentence is to say the sort of thing that Julius Caesar said by *alea jacta est* when he broke the laws of the Roman Republic and, instead of disbanding his troops, led them towards Rome across the boundary marked by the river Rubicon. We who follow Caesar use the English sentence to assert that, in doing some act or other such as crossing that stream, we have committed ourselves irrevocably.

What is the difficulty here? The difficulty this creates for the Frege–Wittgenstein characterization of sense is that it shows that there is no simple route between the ordinary or normal use of a sentence such as "the die is cast" or "alea jacta est" – or from what people usually say by uttering it – and its strictly or narrowly linguistic meaning.

The proper response to this problem is to concede something. We must adjust the Frege–Wittgenstein thesis to read as follows:

> Sentence *s* has as its use in L(i) to say *literally* (to say in the thinnest possible acceptation of 'say') that *p* – thus *s* means that *p* in the narrowest strictest sense of 'means' – if and only if the referential specifications specific to the language L(i) [e.g., the sorts of specification given in §7] rule that whether *s* is true or not depends upon whether or not *p*.

This reformulation simply spells out an intention that Frege or Wittgenstein could have voiced. But what it suggests is that, in order to implement that intention, we have to embed our new formulation in some larger, more comprehensive theory, the sort of theory for which we have to look forward to the work of J. L. Austin.[10] This can persevere in the Fregean explication of the literal meaning of a sentence as consisting in its sense or truth-condition. But the fuller kind of saying that we find in the *the die is cast* example is something that the comprehensive theory will have to explain by building upwards and outwards from literal meaning as characterized after the fashion of provisions like T(1), T(2), P(1), C(1), C(2), and R(1). A neo-Austinian theory may suggest that, by doing the *rhetic act* of uttering something which has as its sense (and means literally) in language L(i) that the die has been thrown, and by performing thus the *locutionary act* of saying 'the die is thrown,' a speaker can perform a further speech act, namely an *illocutionary act*, tantamount in force to the declaration or intimation that he is irrevocably committed. By saying one thing then (here a false thing) Caesar conveys something else, which proves to have been a true thing.

9. There is more to say about this, but it only needs to be shown how one might place in a single focus the Frege–Wittgenstein conception of sense, in the condition in which it

was available by 1921, and the different researches of J. L. Austin. (See also Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING, §3.) These were undertaken some 30 years after the *Tractatus*, in a framework of theoretical expectations both at odds with the concerns of *Grundgesetze* and *Tractatus* and uninformed by attention to very much that these works had in common. But the reason to mention the difficulty here is that, unless we are prepared to use Austin's work to *delimit* the area within which Frege, Russell, and the early Wittgenstein wanted to operate, their theories will be plagued with irrelevant objections. All will be well, however, provided that the theory of literal sense can be fitted into a larger framework that embraces among other things both the non-literal use of declarative sentences and the literal use of ordinary non-declarative sentences.[11]

At this point, let us attend to a passage too rarely heeded as already expressive of Wittgenstein's constant awareness of the importance of such a framework and of the wider frameworks that must contain this one. It occurs at *Tractatus* 4.002:

> Man possesses the ability to construct languages capable of expressing every sense, without having any idea how each word has meaning or what its meaning is – just as people speak without knowing how the individual sounds are produced.
>
> Everyday language is a part of the human organism and is no less complicated than it.
>
> It is not humanly possible to gather immediately from it what the logic of language is.
>
> Language disguises thought. So much so that, from the outward form of clothing it is impossible to infer the form of the thought beneath it, because the outward form of the clothing is not designed to reveal the form of the body, but for different purposes.
>
> The tacit conventions on which the understanding of everyday language depends are enormously complicated.

10. 4.002 presages some of Wittgenstein's later dissatisfaction with his Tractarian philosophy – and anticipates many of the preoccupations of later twentieth-century philosophy. But the thing that must immediately have troubled him about what he had written at 4.024 was the non-operational character of the conceptions of sense and truth that he had espoused in the *Tractatus*. By the time of *Philosophische Bemerkungen*, what he prefers to say is this:

> To understand the sense of a *Satz* means to know how the issue of its truth or falsity is to be decided. (*Philosophische Bemerkungen*, IV. 43). (Wittgenstein, 1975)

This new formulation looks backwards one decade at the doctrine of *Tractatus*, and sideways perhaps at the work of the Dutch mathematician L. E. J. Brouwer. But it is no less recognizable as the antecedent of the infamous claim advanced by the logical positivists of the 1930s – the claim that dominated the 1930s and 1940s and had an even longer period of influence in the philosophy of science – namely that the sense or meaning of a sentence is nothing more nor less than the method of its verification. (For further discussion of this, see Chapter 4, MEANING, USE, VERIFICATION.)

In the next phase of his thinking about linguistic meaning, Wittgenstein came to advance a different claim, namely, that ("for a large class of cases") to understand a linguistic expression is simply to grasp "its use in the language." (See the *Blue* and *Brown Books* and see the two decades' worth of philosophy books by other philosophers who were influenced by this formulation.) As verificationism fell out of favor, this doctrine rushed in to fill the vacuum

that was left by its disappearance.[12] Then, as the limitations came to be perceived of the doctrine of meaning as use, the next conspicuous contribution to the philosophy of meaning was Grice's idea that the meaning of a declarative utterance was a function of speakers' intentions to use that sentence to induce (by the recognition of that intention) this or that belief. (The trouble with that, one might think, was that such an intention was really the intention to *tell* someone something. It seemed to presuppose rather than to explicate the notion of *saying* that thing. It left too much to do under that head.)

The Fregean idea was destined to be rediscovered for philosophy and accorded an attention it had never previously enjoyed, but scarcely immediately.[13] For English speakers, it remained more or less buried until 1959, when Michael Dummett's article "Truth" (1959) disinterred it and put it back into circulation.[14] This limited circulation was yet further narrowed by the fact that Dummett expressed reservations of his own, not dissimilar to those we find in Wittgenstein, about the acceptability of the Fregean equivalence between sense and truth-conditions. (See Chapter 20, REALISM AND ITS OPPOSITIONS, §§1 and 2.)

11. So much then for the shift that Wittgenstein himself seems to have prompted away from the doctrine of *Tractatus* 4.024, and so much for the philosophy of language that had worked itself out over the period between 1921 and the 1960s, downstream of Frege, Russell, and early Wittgenstein, before Davidson's philosophy of language first became visible. But now let us go back to the point in the argument that we had reached at the end of §7.

In §7, having expounded *Grundgesetze* 1.32, we claimed that Frege or Wittgenstein would have been well placed to defend the truth-conditional thesis against the objections mentioned in §4 by formulating it as follows: in *Begriffsschrift* extended (Bg^+), *s* can be used to say literally that *p* if and only if the equivalence [True *s* if and only if *p*] flows from the referential stipulations for the language Bg^+. The difficulty that this left over was this: that the most that this positive doctrine will ever enable us to *put on the page* is an account of what it is for a sentence to say-in-the-language-of-Bg^+ that *p*, or

> *s* can be used to say-literally-in-Bg^+ that *p* – and *s* means-literally-in-Bg^+ that *p* – if and only if it is derivable from the referential-stipulations-for-Bg^+, specified thus …, that *s* is a true-Bg^+-sentence if and only if *p*.

This points at something general, namely the thing that Wittgenstein gets across in 4.024. But how can we articulate this general thing? How can we extricate "mean literally," "say literally," or "referential stipulation" from these hyphenations with "Bg^+"?

12. One manageable objective we might set ourselves is this: to arrive at the generalization we need by satisfying all the necessary conditions to supplant the constant "Bg^+" by a variable "L(i)." If we proceed in this way, how can we make explicit the thing that the Bg^+-relative condition only shows?

Looking back at what we then have to generalize and free from relativity to Bg^+, it will appear that the chief obligation we now incur is to dispense with the reference to particular stipulations such as T(1), T(2), P(1), …, and so on. Instead, we have to say explicitly what sort of thing a referential stipulation is. And perhaps the most natural first suggestion will be that we should advance on the following basis:

> *s* means that *p* in L(i) if and only if there is a Θ for L(i), namely $\Theta^{L(i)}$, that associates each expression of L(i) *with its proper value*, and this $\Theta^{L(i)}$ implies that *s* is true if and only if *p*.

Such a proposal will resonate in multiple ways with a common theme in a variety of semantical traditions. (Davidson calls it the building-block proposal.) The only trouble is that, in practice, it has never been brought convincingly to life. There is nothing both general and foundational to be said, simply in terms of reference, about how "and," "not," "Caesar," and "behind cloud" all have their meaning. We cannot dispense in semantics with something like the idea of reference. Equally, however, we cannot make out of the idea of reference the whole basis for the semantics of the sentence. From a standing start, we cannot even explain in such terms what distinguishes a sentence from a mere list. Frege himself never at any point dispensed with the idea of reference. But he also insisted, in the preface to *The Foundations of Arithmetic*, that "only in the context of a sentence does a word mean or stand for anything." Somewhere near the beginning of our account we have to render it more intelligible than this first suggestion will that sentences can be used not merely to list items of reference but to *say* things. Rather than start with the idea of reference, ought we not try to start at the other end, with truth itself and the *contributions* that the constituents of a sentence make to the conditions for its truth? On these terms, let the meanings of words be just as various as their contributions.

13. Noting that truth and meaning are symmetrically relativized to language in the elucidation of meaning we offered at the end of §11, we shall find a different suggestion we can explore. Not only did $\Theta^{L(1)}$ in §7 state the meanings of each sentence of the language L(1) and the meaning of [each] constituent of L(1). As a by-product of doing that, it fixed systematically and non-accidentally correctly the extension of the predicate "true" as restricted to L(1) sentences. Thus we have it that "'The sun is behind cloud" is true if and only if the sun is behind cloud, "The moon is behind cloud" is true if and only if the moon is behind cloud, and so on. (Such biconditionals are sometimes called *partial definitions* of 'true sentence of L(1).') We need not know which sentences are the true ones or constitute the actual extension of 'true-in-L(1).' But we do have a systematic way to state the principle on which that extension is assembled and, in that however strange or philosophically unwonted sense, we have a 'definition' of 'true-in-L(1).'

So the new thought is this: rather than arrive at an account of meaning by trying to generalize from the idea of designation, why not underwrite the *Tractatus* 4.024 generalization by saying the following?

> for any *s*, *s* can be used to say literally in L(i) that *p* – *s* means literally in L(i) that *p* – if and only if it is derivable from the definition of *true sentence of* L(i) that *s* is true if and only if *p*.

14. Having had recourse, in this last transposition, to the idea of a definition of truth in L(i), the time has come to turn our attention away from the main trend of semantic speculation in analytical philosophy, and away from Jena, Vienna, and Cambridge towards Lwow, Warsaw, and the study that Tarski called the "methodology of the deductive sciences," which was one part of Tarski's contribution to the prewar development of mathematical logic.[15]

The change of orientation is at first surprising. We are inclined but not necessitated in this direction by the formal shape of the problem we have been considering, which relates only to the conceptual lacuna that divides *Grundgesetze* 1.32 from *Tractatus* 4.024. Other directions are thinkable. Yet, given the actual influences that have formed the semantical speculations of nowadays (Davidson's and others), there is no real alternative – however oblique Tarski's concerns are to Davidson's, and however indirect our progress towards a general theory of linguistic meaning may appear.

Let us begin by asking the question how it can have come about, if the theory of Fregean sense was in no way Tarski's preoccupation, that Tarski should have been

interested in identifying a set of axioms for a language L(i) that delivered theorems given in the form [*s* is true in L(i) if and only if *p*]. Why was Tarski interested in axioms delivering the theorems of which philosophers of language such as Davidson and his associates were going to say that they determined the sense or contribution of each of the expressions of L(i)? The answer is that, even though Tarski was not interested in meaning as such, he was interested, and interested in a special way, in truth.[16] He was interested in the idea of truth neither after the fashion of the traditional logic – truth simply as the thing that valid inference preserves – nor after the fashion of philosophers who are exercised by the more mysterious and perennial questions about truth. The sort of thing Tarski was interested in doing was to find ways to compare and contrast the class of true formulas of a given formal language with the class of formulas that the rules and axioms *make provable* there. Embarking on inquiries of this kind, the thing that Tarski needed was a systematic account of what determined the extension of the concept *true*.[17] (Such a systematic account, given in what I have invited the reader to see as a modernization of the method of Frege's *Grundgesetze*, is what he called a 'definition.') But that was not everything he needed. He also needed to find assurance that his account of truth would not be undermined and discredited in the eyes of the community of mathematicians by the ancient paradoxes that exploited that idea, Epimenides's paradox, for instance (cp. Tarski, 1931, p. 110; 1936, p. 252).

Let us take the second of these problems first. Tarski's analysis of the liar paradox and its variants suggested to him that the best way to safeguard the construction he had in mind was to begin with some particular object-language that was itself free from all semantic notions. Once this object-language itself was made determinate, semantic concepts[18] such as satisfaction, truth, and designation,[19] as restricted to that object-language, could be introduced into the metalanguage for that object-language by defining each deliberately, with full formal correctness, in terms drawn from the object-language (or translations of the same into the metalanguage), from elementary set theory and from the formal morphology of the object-language, as given in the metalanguage.[20] On these terms, one could assure oneself that, if the object-language was immune from paradox, then the metalanguage that contained the object-language would be immune too.

15. So far so good. But on what principle was a restricted, paradox-free notion of truth, the concept *true sentence of L(i)*, to be positively characterized? What was the philosophical or intuitive substance of the idea? For his thoughts about this, Tarski turned (by his own account[21]) to his teacher Tadeusz Kotarbinski's book *Elementy Teorji Poznania* (1929), where we find the following passage (itself reminiscent of *Tractatus* 4.061):

> Let us pass to the classical doctrine and ask what is [to be] understood by "[a sentence's or thought's] accordance with reality." The point is not that a true thought should be a good copy or [fac]simile of the thing of which we are thinking, as a printed copy or photograph is. Brief reflection suffices to recognize the metaphorical nature of such a comparison. A different interpretation of "accordance with reality" is required. We shall confine ourselves to the following explanation: "John judges truly if and only if things are thus and so: and things are in fact thus and so."[22]

Spelling out this explanation for the case of some particular sentence, we have

> John judges truly in saying "snow is white" if and only if
> (1) John is right in saying "snow is white" if and only if snow is white
> (2) snow is indeed white.

But then it seems we can have, more simply[23]

> "Snow is white" is true if and only if snow is white.

The chief thing that it seems the definition of "true in L(i)" must do in order to conform to Kotarbinski's requirements is to imply one such equivalence in respect of each sentence of L(i).[24]

But now, having come this far, we shall be moved to ask: how otherwise can the definition of truth in L(i) furnish the thing Kotarbinski required than by doing the sort of thing we have seen that $\Theta^{L(1)}$ did? *This* is how Tarski's path comes to cross the path we have seen Frege's and Wittgenstein's thoughts as marking out. The parties are moving in different directions, but at the intersection there is one common thing each party needs in order to arrive where it is headed. Each party needs to involve itself, for any language that comes into consideration, in something like the exercise conducted in §7.

16. In the light of this, how is the problem to be solved of saying what a referential specification is? Well, if there is this convergence, then Tarski must have the same problem under a different name if he is to say what a definition of truth is. Tarski has to say what such a definition must be like in order to be adequate. The problem is solved as follows:

> A formally correct definition $\Theta^{L(i)}$ of the predicate "true" as applied to L(i) sentences is *materially adequate* if and only if, for every sentence *s* of L(i), Θ implies a biconditional (or so-called T-sentence) in the form [True *s* if and only if *p*], where '*p*' holds a place for a translation of *s* into the metalanguage ML(i).

Tarski calls this provision – which is evidently not itself statable at any level lower than the meta-metalanguage – Convention T.[25] It is simply the generalization of Kotarbinski's desideratum.[26] Similarly, then, each referential specification for L(i) assigns a value to every expression in L(i); and a set of such assignments is materially adequate under the very same condition as Tarski gives. It must yield a T-sentence for each sentence of L(i). And each T-sentence must in the same way be translational. This is to say that, in each case, 'p' must hold a place for a translation of *s* into the metalanguage.

17. Does this represent any progress? For Tarski, it is progress, because Tarski's only objective is to arrive at a non-accidentally and recognizably correct definition of true sentence of L(i). The word "translation" is not being used here in a manner that offends against Tarski's professed attitude to semantic notions. It occurs only in the meta-metalanguage, or (as one might fancifully say) in Kotarbinski's and Tarski's philosophy of truth. Occurring there, it presupposes only this: that a logician can recognize when the sentence given on the right-hand side of a T-equivalence is faithful to the meaning of the sentence mentioned on the left. Nevertheless, because Convention T includes within it a semantical term coordinate with the ideas of *meaning*, *definition*, and the rest, anyone who is concerned with the idea of meaning for its own sake still faces the same old question. How can we eliminate the semantical term "translation" from Convention T? Or how can we analyze or dismantle it there?

Here at last we can resume the story that we have already carried up to 1959, which was the moment, as we saw, when Michael Dummett put the truth-conditional insight back into circulation. If anybody had been concerned with the question of how to make Wittgenstein's generalization 4.024 work, then Tarski's construction would have served him perfectly – unless he had had such an obsessive concern with the nature of meaning itself that it was not sufficient to trace and explore the small circle that joins the ideas of truth,

meaning, and translation. Here though is the trouble, that, perfectly properly, *ex officio*, and by its nature, philosophy is imprisoned within that obsession.

18. To understand Donald Davidson's revival of the general idea of meaning as given by truth-conditions and the distinctive advance that this made possible, it helps to appreciate the immediate background of his speculations. This was not any concern on Davidson's part with the theory common to early Wittgenstein (to whom Davidson rarely, if ever, referred) and Frege (to whose doctrines Davidson evidently regarded Alonzo Church as the complete guide, even though this guide completely omitted all mention of Frege's truth-conditional insight). The background was more topical, namely Davidson's doubts about Carnap's methods of extension and intension,[27] his considered rejection of the answer to the question of linguistic meaning provided by H. P. Grice's reduction of semantic notions to psychological ones such as belief and intention,[28] and Davidson's attachment to the speculative framework furnished by W. V. Quine's book *Word and Object* (1960) – most especially the question of what a thinker from outside a community of speakers would need to avail himself of if he were to try to make sense of utterances in their unknown language. What Davidson wanted was to retain Quine's naturalistic approach to such questions, to align himself with Quine's objection to all "museum myths" of meaning, but to do so without commitment to Quine's talk of ocular irradiation, neural impacts upon subjects, and the rest. According to Davidson, the thing that impinges on subjects had better be the world itself, the world that is common to both interpreter and subjects.

Seeking for some framework within which to give a systematic account of the information (or putative information) that an interpreter would need to amass and draw upon in order to frame his hypotheses about the meanings of his subjects' uttered sentences, and seeking at the same time to sweep away the supposed obscurity of '*s* means that *p*,' the construction Davidson found himself reaching for was in effect none other than Tarski's. Davidson writes:

> Let us try treating the position occupied by '*p*' [in '*s* means that *p*'] extensionally: to implement this, sweep away the obscure 'means that,' provide the sentence that replaces '*p*' with a proper sentential connective, and supply the description that replaces *s* with its own predicate. The plausible result is
>
> (T) *s* is T if and only if *p*.
>
> It is worth emphasizing that the concept of truth played no ostensible role in stating our original problem [the problem of a theory of meaning for a given language]. That problem upon refinement led to the view that an adequate theory of meaning [for the language spoken by the interpreter's subjects] must characterize a predicate meeting certain conditions. It was in the nature of a discovery that such a predicate would apply exactly to the true sentences…. A Tarski-type truth definition supplies all we have asked so far of a theory of meaning. (Davidson, 1967)

The discovery is of course a rediscovery – the rediscovery of the thing that Frege and Wittgenstein had articulated and that Davidson failed to credit to Frege or to Wittgenstein. If Frege's original insight had not been correct, there could have been no such discovery. Working within Quine's framework, however, the attitude Davidson had towards Tarski was as follows. Taking translation for granted (or taking "means in L(i)" for granted), Tarski had defined "true sentence of L(i)." Conversely, then, why should not Davidson take truth in L(i) for granted, in order to define "means in L(i)"? The residual problem was then to dispense with Tarski's use of the word "translation."

19. Davidson's first thought about that problem seems to have been that he could secure everything he needed if he were simply to omit the requirement that the T-sentences generated by $\Theta^{L(i)}$ in the form [*s* is true if and only if *p*] should provide translations on the right-hand side of the L(i) sentence *s* mentioned on the left. Could he not stipulate instead that absolutely all the T-sentences that $\Theta^{L(i)}$ generated should be true? But it is now pretty clear that the condition is not sufficient.[29]

From the beginning of all Davidson's speculations, however, shaped as they were by Quine's *Word and Object*, the real solution to this problem was always at hand. Perhaps Davidson's best account of this solution is the one given in his "Radical interpretation" (1973).[30] But there is a real point here in giving a Davidsonian solution in a variant that is not open to the objections that so many critics have urged against the particularities of Davidson's own formulation.[31] The distinctive features of the variant presentation are chiefly due to Richard Grandy and John McDowell.[32]

If the interpreter of the utterance of a sentence is to say what it means, then he has to find out under what conditions the sentence, being the sentence it is, counts as true. To say so much is to say little more than Frege said. But the next thought one will have reaches beyond Frege. It is that linguistic behavior is a proper part of behavior. But, if so, there ought to be some other than purely simply semantic way of specifying what it is for a radical interpreter to succeed in interpreting alien subjects. Surely the interpreter's linguistic efforts are part of the larger effort to interact successfully with subjects, to coordinate his/her (the interpreter's) practical doings with those of subjects … in a word, to make sense of subjects and have them make sense of him/her. If we enlarge in this way, in terms which are not *specifically* semantical, upon what interpretation is attempting to achieve – and if we count the interpretation of speech as one proper part of the larger undertaking – then here at last we arrive at the substantive non-semantic constraint upon $\Theta^{L(i)}$ that we have been looking for:

> A definition of truth in L(i) will be materially adequate if it generates a T-sentence for each sentence *s* of L(i) and collectively the T-sentences that the definition implies, when experimentally applied to individual utterance by the speakers of L(i), advance unimprovably the effort to make total sense of the speakers of L(i).

The notion of total sense is not a semantic notion, but it *subsumes* one. One person's making sense of another is a matter of their participative interaction in a shareable form of life, of their homing upon the same objects, of their being in a position *ceteris paribus* to succeed in joint enterprises, and so on. In so far as we make sense of others, we deploy a mode of understanding that can be redescribed, however artificially, as follows. There is a store of everyday predicates of human subjects, of features of the environments that impinge on subjects, and of the events that are counted as the actions or conduct of such subjects. When we seek to attain participative understanding, we seek in response to circumstances, including the speech or conduct of subjects, to distribute predicates of these and other kinds across features of reality, mental states, and actions in such a way that: (1) the propositional attitudes we ascribe to subjects, specifying the content of these attitudes, are intelligible singly and jointly in the light of the reality to which we take subjects (or their informants, or their informants' informants …) to have been exposed; and (2) the actions (and actions of speaking) that we ascribe to subjects are intelligible in the light of the propositional attitudes we ascribe to them.[33]

In the form in which we now have it, the new elucidation of meaning finally bridges the gap between Frege's stipulations for his concept-script and Wittgenstein's bold generalization

of Frege's idea. Of course, it inherits all the well-known difficulties of the ideas of understanding, explaining, making intelligible, imaginative projection, or identification. But these difficulties are there anyway. The proposal not only depends upon these ideas. It assists us by helping to trace their interrelations.

20. The conclusion to which we have been drawn is that what it is for a sentence to mean that the sun is behind cloud and to be available to say that the sun is behind cloud, is as complicated as this. It involves a biconditional, "'The sun is behind cloud" is true if and only if the sun is behind cloud,' which is imbedded within the scope of an operator whose presence indicates that this biconditional is derivable from the whole system by which we make sense to one another and make sense of one another. What we have here is the idea of a significant language as a system that correlates strings of repeatable expressions with the states of affairs that the strings can draw attention to or get across, this system itself being a subsystem of the larger system by which social beings participate in their shared life. There is nothing abstruse in that. It is because we grasp it so readily (I think), both in philosophy and before philosophy, that we can hear a T-sentence given in the form "*s* is true if and only if *p*" as the output of such a system. When we grasp that, it is tantamount to our grasping something intuitively similar to the $\vdash_{\Theta(\mathrm{Bg}+)}$ that played the part we described in the Fregean elucidation of the meaning of *Begriffsschrift* extended sentences.

21. Objection may be made because, in the formulation I have set down here, *s* can only have it as its literal use to say that *p* if *all* suitably constrained theories imply that *s* is true if and only if *p*. What reason is there to suppose that this condition is non-vacuously satisfiable? The objection is a good one, because the formulation does seem to foreclose a matter that ought to have been left open. It seems better on reflection to postpone such questions until we have a fuller account of what it is to make sense of the shared life and conduct of L-speakers. This is a question of the indeterminacy of interpretation – or of translation, as Quine says. (See Chapter 26, INDETERMINACY OF TRANSLATION.) In the interim, perhaps we should rule that it is sufficient for *s* to mean that *p* that *some* putatively unimprovable theory that meets all the constraints should entail the biconditional [*s* is true if and only if *p*].

22. It may be objected that the idea of translation that our final proposal purported to remove surreptitiously returns with the idea of an interpreter's 'making sense' of other people. But to this the theorist of truth-conditions must reply by simply reiterating his claim that the idea of making sense that we find here is a much wider one than the idea of linguistic interpretation. The presence or absence of this more general thing can be demonstrated non-linguistically. The ideas of making sense of and being made sense of embrace and subsume the ideas of saying and the interpretation of saying, and they involve them illuminatingly with coeval, collateral ideas of explanation and understanding – even (as you may say, if you are as convinced as I am of the indispensability of these further things to the full story) with the idea of participation by interpreter and subjects in a shared form of life, and the idea of explanation as *Verstehen*.

23. A third objection might take the following form. After all the changes and emendations consequential upon earlier objections, should not all residues of the idea of compositionality itself have been expelled from the final formulation? "Truth itself is unduly emphasized in your construction," the objector may say. "One might accept this for argument's sake as the result of your foolish concentration upon declarative utterances. But, even in the cases where truth really does belong, it is surely not necessary to insist that the interpretive biconditional should be generated by the recursively or compositionally generated definition of truth that you envisage for the language L(i). If we are simply helping ourselves

now to the idea of what it requires to "make total sense" of speakers, *Verstehen* and the rest, why cling to this residue of Fregean compositionality?"

To this I would reply that the meaning we are interested in understanding is linguistic meaning, the non-natural meaning possessed by sentences that will be further saturated by context of utterance (etc.). It is the meaning with which sentences of what we recognize as languages are invested. (See Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING, §5ff.) Generally speaking, what makes interpretation possible is the fact that the language to which the sentences belong can be treated as *pre-existing* any particular speaker or hearer and any particular act of communication. It is something that speakers and hearers need to know about already. The compositionality that theories of L(i)-sense or definitions of 'true sentence of L(i)' have to reflect is a property of the language L(i) itself, L(i) and its properties being something irreducible to any psychological, social, or pre-linguistic fact or facts about individual speakers or individual situations of communication.[34]

24. In opposition to such claims as the one just entered, many have tried to see the clauses of the definition of 'true sentence of L(i)' as answerable, in the last analysis, to psychological or neurolinguistic facts about speakers. After further reflection, some among those who are tempted by such an approach have shied away from the manifest embarrassments of getting involved in all that. And, backing off, they have preferred to say (as John Foster and Donald Davidson have more or less agreed in saying[35]) that the "theory" corresponding to the definition of true sentence of L(i) "explicitly states something knowledge of which *would suffice* for interpreting utterances by speakers of the language."[36] There are doubts about this kind of formulation. My own view would be that the question it answers should never have been permitted to arrive at the point where it could exact either this or any remotely similar answer. The thing the definition of truth for L(i) is answerable to is how things are with the social object that is the language L(i) – not how things are with the speakers past and present in virtue of whose existence that language is extant. The question for anyone who would define truth for L(i) is this: how have we to see L(i) – how must we parse it and segment it – in order to understand why its sentences mean this or that? How do we have to see L(i) in order to get principles by which we work out what its more complicated or obscure L(i) utterances mean? Again, why do L(i) sentences have to be translated into foreign tongues on *this* principle rather than that principle in order to arrive at a passable version of what was originally said? In so far as purposes such as working out what sentences mean and discerning principles of translation do not force us into one sort of grammatical description rather than another, there may be indeterminacy about the properties of L(i). But that is nothing new. Nor does it render it indeterminate which object the language L(i) is. L(i) is a historically given thing, changeable no doubt, and always in process, but a persisting social object nevertheless (see here Wiggins, 1997). It is not in any reprehensible sense an indeterminate or mythical object.

25. One last question. What, then, after all these twists and turns, was the advantage of going by the Tarskian route to our final destination? One alternative might have been to reflect that we never really define or reduce anything in philosophy. So someone might ask: Why not gloss the notion of meaning in a freewheeling fashion by simply using it and involving it with all the collateral notions that are imported by the idea of interpretation?[37] Such, after all, is the method of philosophical elucidation – the method we have learned not to hope to improve upon.

There is much to agree with in this objection – the Davidsonian account is an exercise in elucidation too – except that the one principal contention seems wrong. It seems wrong to

suggest that we should deny truth its foundational place in the elucidation of meaning. For there is a real advantage in going by the Fregean and Tarskian way. It is true that Tarski's construction, which consolidates Frege's, is conditioned in the first instance by Tarski's deep suspicion of primitive semantic notions, and this is a suspicion one may not share. But suspicion of the semantical as such is not the only possible reason one might have to applaud the fact that Tarski gives his construction in terms of simple truth (*not* truth in a structure/model),[38] that he introduces semantic notions deliberately and in a measured fashion, and defines notions like satisfaction and the valuation function (*) by fixing their extension. One may applaud all this not because one thinks semantic notions really *are* suspect, but because an account of meaning that builds on Tarski's construction *helps to show how meaning* is *possible*. By seeing the definition of 'true sentence of L(i),' for any language L(i) as needing to be built up in this careful and austere fashion, while the output of the definition is constrained in a manner that is irreducibly non-austere (as messy and anarchical as the social always will be), we can understand something about how it is possible for there to be such a thing as the semantical, and on what conditions it is possible, namely the existence of both the compositional (in the small) *and* the social (in the large).

Notes

1 "On the object of my concept-writing" (1883) in *Nachgelassene Schriften*, translated in *Posthumous Writings*. For the concept-writing itself, see *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (1879).

2 See *Grundgesetze der Arithmetik* (1893). In this work, §32 and the preceding sections consolidate, codify, and complete the doctrines of (direct) sense and reference explored and expounded in "Ueber Sinn und Bedeutung," pp. 25–50.

3 See *Begriffsschrift* (1879, n. 1).

4 In view of the confusion surrounding this mathematical term, Frege did not call them 'variables.'

5 *Tractatus Logico-Philosophicus* (1921). I translate *Satz* here not as 'proposition' but as 'sentence in use,' in order to mark and preserve the continuity (as well as the discontinuity) with Frege, who always used *Satz* to mean what we now mean by 'sentence.' I think Wittgenstein effectively answers the complaint that Frege has nothing to say about what it is to understand a sentence or grasp a thought. For this complaint – justifiable enough, perhaps, when directed against such traditional accounts as the one given in Church (see §04 of *Introduction to Mathematical Logic*, 1956) – see, e.g., Dennett, *The Intentional Stance* (1987, p. 123): "Frege does not tell us anything about what grasping a thought consists in …" In fact, it would be much fairer to complain against him (if one thinks this a matter for complaint) that, by introducing the thought as that which one grasps by virtue of grasping the acceptance/rejection conditions of something linguistic, Frege may seem poised to acquiesce, not in a vacuous Platonism of *noeta*, but in a potentially highly controversial quasi-linguistic view of thinking as the soul's internal dialogue with itself. Interestingly, this view really is Platonic: "The soul when it thinks is simply conversing with itself, asking itself questions and answering, affirming and denying…. So I define one's thinking as one's speaking – and one's thought as speech that one has had – not with someone else or aloud but in silence with oneself" (Plato, *Theaetetus* 189^{E}–190^{A}). On this and cognate matters, see now Dummett, "The philosophy of thought and the philosophy of language" (1986).

6 For Davidson's version, see "Truth and meaning" (1967). And see below §18. For various formulations of the difficulty explained in the text above and cognate apparent difficulties, see Ayer, "Truth" (1953); Wiggins, "On sentence-sense, word-sense, and difference of word-sense," (1971, pp. 18–19); Strawson, "Meaning and truth," inaugural lecture (1969); and Foster, "Meaning and truth theory" (1976). See also Davidson's "Reply to Foster" (1976), on which see below, n. 29.

7 Here I borrow an expository idea from Michael Dummett. See his *Frege: Philosophy of Language*, pp. 227–228.

8 For some discussion of these issues, see my "On the sense and reference of predicate expressions," with references there to V. Dudman and P. Sen.

9 For the use of the relative pronoun 'what' in connection with the references of predicates, see Frege, *Posthumous Writings*, p. 122. See also Dummett, *Frege: Philosophy of Language*, pp. 211–217.

10 For J. L. Austin's theory of locutionary, illocutionary, and perlocutionary acts, see *How to Do Things with Words* (1962). A rhetic act is an act of using vocables with a contextually determinate sense and reference and in such a way that one can be reported as saying that … For the connection between the locutionary and the rhetic, for the connection between Austin's researches and post-Austinian developments, and for much else besides that belongs in the areas I have so roughly blocked in, see Hornsby, "Things done with words."

11 If the inner core of a theory of sense for a given language is stated truth-conditionally, then the immediately adjacent next outer portion of that larger theory comprises the theory of the other linguistic moods of L(i). This will identify linguistic acts as acts of specifically *asserting that* [the sun is behind cloud, say], *asking whether* [the sun is behind cloud], or *enjoining* (again in the thinnest possible sense, and however vaingloriously in this particular case) *that* [the sun be behind cloud]. Cp. McDowell, "Truth conditions, bivalence and verificationism," p. 44, who assigns this task to a "theory of force." For the reasons why one might hive this task off from a theory of force in Austin's more general sense, see Davidson, "Moods and performances," pp. 109–121. See also Hornsby, "Things done with words."

12 My recollection from being an undergraduate at Oxford during the 1950s at the time when Austin was giving the lectures he then called *Words and Deeds* (1954–1955), but before the appearance of Grice's article "Meaning" (1957), is that in that period the doctrine then current about the meaning of words and sentences was simply a generalization of the Wittgensteinian thesis that meaning was use. There was no audible trace of the idea that to know the meaning of a sentence was to know what it would take for it to be true. To judge by my experience three years later in the Princeton philosophy department, the situation was very much the same in North America.

13 It is true that in the 1950s Frege's writings were being translated. But neither *The Foundations of Arithmetic* nor "On sense and reference" (the one paper which Carnap, Quine, Feigl, and Sellars had made familiar to all professional philosophers) explained what the sense of a sentence was to be. Nor did any of Geach's and Black's other *Selections*. It is true, too, that *Tractatus* 4.024 was legible enough. But, by its apparent archaism, the picture theoretical-framework obscured the doctrine.

14 Dummett, "Truth." It is noteworthy that in the several decades here under consideration, Wittgenstein's is the one clear philosophically salient formulation of the connection that Frege discerned between sense and truth-condition. Frege's doctrine on this point is conspicuous by its absence from expositions where we might have expected to find it, such as those of Alonzo Church at §04 of his introduction to *Introduction to Logic* (1956) and Rudolf Carnap at §33 of *Der Logische Aujbau der Welt* (1928). (For Carnap's own insufficiently remarked final return to a Fregean position, without explicit acknowledgment to Frege, see *Introduction to Semantics*, 1944, p. 22.)

I have wondered whether it is something connected with the blind spot I seek to explain in the text that accounts for the strange neglect of Richard L. Cartwright's definitive improvement (1954) of Quine's criterion of ontological commitment, namely his reformulation of this in terms of rules of truth. See Cartwright's "Ontology and the theory of meaning," an article that rehearses and resolves difficulties that were still under active discussion a whole decade later.

15 By "methodology of the deductive sciences" was meant, *inter alia*, the systematic study of such notions as *sentence*, *consequence*, *definition*, *deductive system*, *equivalence*, *axiom system independence*, *consistency*, and *completeness*.

16 See Tarski, "The concept of truth in formalized languages"; also "The semantic conception of truth" and "Truth and proof."

17 Having determined the extension of these concepts, of the true and the provable, he could then inquire whether they coincided. Tarski went on to show that the metalinguistic definition of 'provable in L(i)' – a purely syntactical notion – could be given within L(i); but that, for any L(i) of sufficient expressive power, the semantic paradoxes would obstruct the definition of 'true in L(i)' in L(i).

18 That is, as Tarski puts it, "concepts which, roughly speaking, express certain connections between the expressions of a language and the objects and states of affairs referred to by those expressions."

19 The extensionally defined counterpart of reference is the valuation or asterisk function as it is defined for each L(i). For the importance of not *beginning* by calling this function that of 'reference,' see McDowell, "Physicalism and primitive denotation."

20 The metalanguage is the language in which one may speak of whatever the object-language speaks of and also of the expressions of the object-language in their relation to that which the object-language speaks of.

21 See the Bibliography to Tarski, "The concept of truth."

22 *Elementy Teorji Poznania*, pp. 106–107 in the English translation. Note that neither Kotarbinski nor Tarski takes this schema to be the recipe for a redundancy, deflationist, or (as Tarski says) nihilistic theory of truth. Indeed, Tarski sometimes claimed to be coming to the rescue of the correspondence theory – though this claim must be taken with a pinch of salt. (Nothing in Tarski's theory can vindicate the idea that truth is to be defined in terms of a relation between sentences and states of affairs. Nor is there anything essential to the Tarskian construction that will vindicate the classical conception of truth as bivalent. Such questions remain open.)

23 For the claim about Tarski and Kotarbinski, see Wiggins, *Needs, Values, Truth*, pp. 333–334. (In addition to making general reference to Kotarbinski's book, Tarski refers also to lectures in Warsaw by Lesniewski. But the main burden of that acknowledgment seems to relate to the semantic paradoxes.)

24 For the failure of proposals to deliver this result by the method (which is not Tarski's official method) of simply conjoining 'partial definitions,' see Milne (1999).

25 Cp. "The concept of truth," p. 187. The words in the text above are an application of Tarski's doctrine, not a quotation.

26 Material adequacy is adequacy to the subject-matter, which is truth. It therefore entails *non-accidental* fidelity to the extension of the predicate. To think here of the material conditional/biconditional will excite precisely the wrong associations. See Tarski (1931, p. 129).

27 Davidson, "Carnap's methods of intension and extension."

28 See again Grice, "Meaning." See also Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING.

29 It can be proved that, if there is one theory that provides a true T-sentence for each sentence of the language L(i), then there will automatically be a second such theory, and the interpretations to be read off the second theory will be different from those to be read off the first. See Evans and McDowell's editorial introduction to *Truth and Meaning* (1976). Their finding is not superseded by the footnote that Davidson added in 1982 to the *Inquiries* reprint of "Truth and meaning" (p. 26, n. 10) however illuminating the footnote might be in other ways.

30 See also Chapter 13, RADICAL INTERPRETATION.

31 Objections have mostly related to Davidson's free-wheeling use of the idea of an interpreter's needing to find what sentences a subject *holds true*. It must be noted, however, that Davidson has persisted in this part of his original presentation, and has developed it further in his Dewey lectures, *Journal of Philosophy* (1990).

32 See Grandy, "Reference, meaning and belief"; Evans and McDowell, editorial introduction to *Truth and Meaning*; McDowell, "Truth conditions, bivalence and verificationism," §1; and McDowell, "On the sense and reference of a proper name."

33 See McDowell, "On the sense and reference of a proper name."; also for some further suggestions, see Wiggins, *Sameness and Substance*, p. 222, and *Needs, Values, Truth*, ch. 4 (*ad init.*).

34 That is to say that I propose that one see language as a social object with a past, a present, and a future, something that is for each generation of speakers an *objet trouvé*, with words and modes of combination possessed contingently of this, that, or the other meaning. Languages are not, on this conception, abstract objects defined by their syntax or semantics. (As Nietzsche remarks, nothing with a history can be defined.) What the syntax and semantics (as of *t*) are answerable to is the state of this language (as of *t*), not the states of the speakers who aspire to speak that language.

35 See their respective contributions to Evans and McDowell, *Truth and Meaning*.

36 That is to say that they shy away from claiming that this is the theory that speakers actually use. Davidson, however (who has so much to lose from misunderstanding here), has not, when he has spoken of speakers and interpreter's 'theories,' exercised all the caution I should have counseled on this matter. See, for one instance among several, "A nice derangement of epitaphs."

37 See, e.g., the approach to meaning of Sainsbury, "Understanding and theories of meaning," pp. 127–144; and of Davies, *Meaning, Quantification, Necessity*.

38 On this point, see again Milne (1999).

References

Austin, J. L. 1962. *How to Do Things with Words*, edited by J. O. Urmson. Oxford: Clarendon Press.

Ayer, A. J. 1953. "Truth." *Revue Internationale de Philosophie*, 25: 183–200.

Carnap, R. 1928. *Der Logische Aufbau der Welt*. Berlin: Weltkreis-Verlag.

Carnap, R. 1944. *Introduction to Semantics*. Cambridge, MA: Harvard University Press.

Cartwright, R. L. 1954. "Ontology and the theory of meaning." *Philosophy of Science*, 21(4): 316–325.

Church, A. 1956. *Introduction to Mathematical Logic*. Princeton, NJ: Princeton University Press.

Davidson, D. 1963. "Carnap's methods of intension and extension." In *The Philosophy of Rudolf Carnap*, edited by P. A. Schilpp. La Salle, IL, and London: Cambridge University Press.

Davidson, D. 1967. "Truth and meaning." *Synthese*, 17: 304–323. Reprinted in *Inquiries into Truth and Interpretation*, 1984. Oxford: Oxford University Press.

Davidson, D. 1973. "Radical interpretation." *Dialectica*, 27: 313–328. Reprinted in *Inquiries into Truth and Interpretation*, 1984. Oxford: Oxford University Press.

Davidson, D. 1976. "Reply to Foster." In Evans and McDowell, 1976, pp. 33–41.

Davidson, D. 1979. "Moods and performances." In *Meaning and Use*, edited by A. Margalit. Dordrecht, Netherlands: Reidel. Reprinted in *Inquiries into Truth and Interpretation*, 1984. Oxford: Oxford University Press.

Davidson, D. 1986. "A nice derangement of epitaphs." In *Truth and Interpretation*, edited by E. Le Pore. Oxford: Blackwell.

Davidson, D. 1990. "The structure and content of truth." *Journal of Philosophy*, 87(6): 279–328.

Davies, M. K. 1981. *Meaning, Quantification, Necessity*. London: Routledge.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: Bradford Books.

Dummett, M. 1959. "Truth." *Proceedings of the Aristotelian Society*, 59: 141–162.

Dummett, M. 1973. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. 1986. "The philosophy of thought and the philosophy of language." In *Mérites et limites des méthodes logiques en philosophie*. Paris: Vrin et Fondation Singer Polignac.

Evans, G., and J. McDowell, eds. 1976. *Truth and Meaning: Essays in Semantics*. Oxford: Oxford University Press.

Foster, J. 1976. "Meaning and truth theory." In Evans and McDowell, 1976, pp. 1–32.

Frege, G. 1879. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle: Verlag von Louis Nebert. Republished and translated in *Conceptual Notation and Related Articles*, edited by Terrell Ward Bynum. Oxford: Clarendon Press, 1972.

Frege, G. 1883. "On the object of my concept-writing." In *Nachgelassene Schriften*. Republished in *Posthumous Writings*, edited by H. Hermes, F. Kambartel, and F. Kaulbach, translated by P. Long and R. White. Oxford: Blackwell, 1979.

Frege, G. 1884. *Die Grundlagen der Arithmetik, eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Hamburg: F. Meiner. Republished in *The Foundations of Arithmetic*, translated by J. L. Austin. Oxford: Blackwell, 1950.

Frege, G. 1892. "Ueber Sinn und Bedeutung." *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50. Republished and translated in *Readings in Philosophical Analysis*, edited by H. Feigl and W. Sellars. New York: Appleton-Century-Crofts, 1949. Also republished in *Translations from the Philosophical Writings of Gottlob Frege*, edited and translated by P. T. Geach and M. Black. Oxford: Blackwell, 1952.

Frege, G. 1893. *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet*, vol. 1. Jena, Germany: Verlag Hermann Pohle.

Grandy, R. 1973. "Reference, meaning and belief." *Journal of Philosophy*, 70(14): 439–452.

Grice, H. P. 1957. "Meaning." *Philosophical Review*, 66(3): 377–388.

Hornsby, J. 1988. "Things done with words." In *Human Agency: Language and Duty: essays for J. O. Urmson*, edited by Jonathan Dancy, J. M. E. Moravcsik, and C. C. W. Taylor. Stanford, CA: Stanford University Press.

Kotarbinski, T. 1929. *Elementy Teorji Poznania. Republished in 1966 in Gnosiology: the scientific approach to the theory of knowledge*, edited by G. Bidwell and C. Pinder, translated from the 2nd Polish edn by Olgierd Wojasiewicz. Oxford and New York: Pergamon Press.

McDowell, J. 1976. "Truth conditions, bivalence and verificationism." In Evans and McDowell, 1976, pp. 42–66.

McDowell, J. 1977. "On the sense and reference of a proper name." *Mind*, 86: 159–185.

McDowell, J. 1978. "Physicalism and primitive denotation." *Erkenntnis*, 13: 131–152.

Milne, P. 1999. "Tarski, truth and model theory." *Proceedings of the Aristotelian Society*, 99: 141–167.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Sainsbury, R. M. 1979/1980. "Understanding and theories of meaning." *Proceedings of the Aristotelian Society*, 80: 127–144.

Strawson, P. F. 1971. *Logico-Linguistic Papers*. London: Methuen. From "Meaning and truth," inaugural lecture, Oxford, 1969.

Tarski, A. 1931. "On definable sets of real numbers." In *Logic, Semantics, and Metamathematics*, translated by J. H. Woodger. Oxford: Oxford University Press, 1955.

Tarski, A. 1936. "The concept of truth in formalized languages." In *Logic, Semantics, and Metamathematics*, translated by J. H. Woodger. Oxford: Oxford University Press, 1955.

Tarski, A. 1944. "The semantic conception of truth." *Philosophy and Phenomenological Research*, 4(3): 341–376.

Tarski, A. 1967. "Truth and proof." *Scientific American*, 220: 63–77.

Wiggins, D. 1971. "On sentence-sense, word-sense, and difference of word-sense." In *Semantics: An Interdisciplinary Reader*, edited by D. D. Steinberg and L. A. Jacobovits, pp. 14–34. Cambridge: Cambridge University Press.

Wiggins, D. 1980. *Sameness and Substance*. Oxford: Blackwell.

Wiggins, D. 1984. "On the sense and reference of predicate expressions." *Philosophical Quarterly*, 34(136): 311–328.

Wiggins, D. 1991. *Needs, Values, Truth*, 2nd edn. Oxford: Blackwell.

Wiggins, D. 1997. "Languages as social objects." *Philosophy*, 72(282): 499–534.

Wittgenstein, L. 1921. "Logisch-philosophische Abhandlung." In *Annalen der Naturphilosophie*. Republished and translated as *Tractatus Logico-Philosophicus*, 1922.

Wittgenstein, L. 1958. *Blue and Brown Books*. Oxford: Blackwell; New York and London: Kegan Paul, Trench, Trubner.

Wittgenstein, L. 1975. *Philosophical Remarks*, edited by R. Rhees, translated by R. Hargreaves and R. White. Oxford: Blackwell.

Further Reading

Davidson has not only proposed an interpretive cum truth-conditional understanding of declarative meaning that inherits the role of Frege's account of these matters – on this consult his (1967), (1973), and (1976) – and made important suggestions about how we should bring the meanings of declarative and other utterances into a general framework – see his (1979). He has also made detailed proposals about the framing of truth-definitions for sub-languages of English that exemplify the difficulties posed by particular modes of combination, most notably the constructions involving reported speech and adverbial qualification. It will be instructive for anyone with an interest in the truth-conditional conception to consult some or all of his papers: "The logical form of action sentences," "On saying that," "Theories of meaning and learnable languages" (all in his *Inquiries into Truth and Interpretation*), and "Adverbs of action," in B. Vermazen and M. B. Hintikka (eds). *Essays on Davidson: Actions and Events* (Oxford: Oxford University Press, 1985).

In connection with these problems as well as with the broadly Davidsonian or neo-Fregean approach to meaning, the reader should study Evans and McDowell's editorial preface to *Truth and Meaning*, Evans and McDowell (eds) (1976); E. LePore (ed.), *Truth and Interpretation* (Oxford: Oxford University Press, 1986); Wiggins "'Most' and 'all': some comments on a familiar programme and on the logical form of quantified sentences," in M. Platts (ed.), *Reference, Truth and Reality* (London: Routledge and Kegan Paul, 1980); McDowell (1978).

For a better understanding of truth in general, see Tarski's (1967) *Scientific American* article, and his (1936) paper, up to, say, definition 23. For a textbook account of truth, satisfaction, and the modern idea of truth in an interpretation, see E. Mendelson's *Introduction to Mathematical Logic* (Princeton: van Nostrand, 1964), pp. 50–53.

3

Intention and Convention in the Theory of Meaning

STEPHEN SCHIFFER

What is the relation between language and thought – or, more exactly, between the representational, or intentional, characteristics of language and those of propositional attitudes such as believing and intending? An answer many find attractive is that we must "explicate the intentional characteristics of language by reference to believing and to other psychological attitudes" (Chisholm, 1958, cited in Speaks, 2010). In other words, it is only our intentional mental states – believing, intending, and the like – that have *original* intentionality, intentionality that doesn't have its source in something else's intentionality; the intentionality of words and speech acts is *derived* intentionality, intentionality inherited from that of associated mental states. That answer raises two questions: how is the original intentionality of thought to be explained, and how does the intentionality of language "derive" from the original intentionality of thought? The focus of this chapter will be on the second question, although the first question can't be ignored entirely, and I shall touch on it in the final section of this chapter.

Although many philosophers seem to accept the derivation view, hardly any of them attempt to spell out how it works. In fact, I can think of only two programs that attempt to explain how the intentionality of language reduces to that of thought, and only one of these programs ventures to reduce all questions about linguistic representation to questions about mental representation. The more ambitious program is one that derives from the Grice-inspired program of intention-based semantics (IBS); the other derives from David Lewis's project of defining what I shall call the *public-language relation*. I shall start with a reconstruction of Lewis's account of the relation in *Convention* (Lewis, 1969) because a problem that immediately arises for that account provides a natural segue to the more ambitious IBS project.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

1 Lewis on the Public-Language Relation

Hardly any philosopher of language would deny that if something is an expression which has meaning in a population, then that is by virtue of facts about the linguistic behavior and psychological states of members of that population. Philosophers of language would like to know which facts those are. They would also like to know what it is for something to be the language of a population. The two questions are apt to seem very closely related, for it's apt to seem that an expression has a certain meaning in a population just in case it has that meaning in the language of that population. If one both thinks of a language as a "pairing of sound and meaning over an infinite domain" (Chomsky, 2006, p. 92) and accepts that an expression *e* has meaning just in case there is something *x* such that *e* means *x*, then one may think of a language as a function – doubtless a finitely specifiable function – that maps finite sequences of sounds or marks (or whatever) onto meanings (or, to accommodate ambiguity, sets of meanings). Then, if *L* is such a function, we may say, first, that

$$e \text{ means}^* x \text{ in } L \text{ iff } L(e) = x$$

(where '*' is to remind us that this stipulated sense of 'means*' isn't the use-dependent notion of meaning that philosophers struggle to understand), and then, second, that

$$e \text{ means } x \text{ in } P \text{ iff, for some } L, L \text{ is a language of } P \text{ and } e \text{ means}^* x \text{ in } L.$$

A language, thus conceived, pairs the words, phrases, and sentences of a language onto the things they mean* in the language. That is bound to seem inevitable, for isn't the meaning of a sentence determined by the meanings of its constituent morphemes in conjunction with the semantic import of the syntactic structures deployed in the sentence? And doesn't that thought lead inexorably to the thought that an account of expression-meaning must first say what it is for a morpheme to have a certain meaning and what it is for a syntactic structure to have a certain semantic import, and to do that in a way that will determine a meaning for every expression of the language? If that is one's sense of how things must be, then the project of saying what it is for a function of the kind just described to be the language of a population ought to strike one as bewilderingly complex. Where is one even to begin, and how could one possibly complete the project without being able, first, to specify a generative grammar for the language, where that is a finitely specifiable theory of the language that generates one or more syntactic structures for each expression of the language and interprets those structures both phonologically and semantically, and then, second, to specify the myriad interlocking practices whereby the morphemes and structures of the language would come to mean in the population what by stipulation they mean* in the language? Anyone who has banged her head against the apparent inevitability, and then the apparent impossibility, of that approach must appreciate the genius and simplicity of David Lewis's way of cutting through that Gordian knot.

In presenting the views I will discuss in this chapter my focus will be on their essential plot lines, and I will simplify like crazy in order not to get bogged down at every turn with technical complexities. To that end I will pretend that the languages we speak have no indexicality, ambiguity, or moods other than the indicative, and that the meaning of a sentence is a proposition, in the generic sense of an abstract, mind- and language-independent entity that has a truth-condition, which it has both necessarily and absolutely (i.e., without

relativization to anything else; see Schiffer, 2003, ch. 1). Relative to those simplifications, I will say that a *Lewis-language* is any function L from finite sequences of sounds or marks (or whatever) – the "sentences" of L – onto propositions. This allows us to say not only that a sequence σ means* the proposition q in the Lewis-language L just in case $L(\sigma) = q$, but also that:

> For any Lewis-language L and sequence of sounds σ, σ is *true* in L iff for some q, σ means* q in L and q is true.

Now a language may be used in any number of ways. For example, a language whose "sentences" are sequences of neural activity may function as a person's language of thought. David Lewis's interest in *Convention* is in a population's using a language as a public language of communication, a language they use to communicate with one another, and his book aims to say what relation must hold between a Lewis-language L and a population P in order for it to be the case that P uses L as a public language of communication. Let's call that relation, whatever it turns out to be, the *public-language relation*. If L is the language members of P use as their medium of communication, then every sentence of L will mean in P what it means* in L. Since the meaning a sentence has in a population is the use-dependent notion of sentence-meaning philosophers want to understand, an account of the public-language relation would be an account of the use on which a sentence's meaning depends, provided that the languages we speak really are Lewis-languages. If they are, and if the only intentionality involved in the account of the public-language relation is that of propositional-attitudes, then the account would have succeeded in defining the intentionality of sentences in terms of the intentionality of thought.

Lewis's definition of the public-language relation in *Convention* is the following, minus an addendum which I will get to presently:

> For any Lewis-language L and population P, L is a public language of P iff there prevails in P a convention of truthfulness in L.

Roughly speaking, a convention for Lewis is a regularity in behavior to which the members of a population want to conform if (nearly) everyone else in the population conforms, and to which they do conform because it's common knowledge among them that they expect one another to conform. Such regularities are self-perpetuating in that past conformity gives rise to the expectation of conformity, which gives rise to future conformity.

In *Convention*, Lewis said that

> It is *common knowledge* in a population P that ___ if and only if some state of affairs A holds such that:
>
> (1) Everyone in P has reason to believe that A holds.
> (2) A indicates to everyone in P that everyone in P has reason to believe that A holds.
> (3) A indicates to everyone in P that ___. (Lewis, 1969, p. 56)

I independently introduced a similar notion in *Meaning* (Schiffer, 1972), which I called *mutual knowledge*,[1] and said that x and y mutually know q just in case x knows q, y knows q, x knows that y knows q, y knows that x knows q, x knows that y knows that x knows q, and so on, and I proposed finite conditions for the generation of mutual knowledge. The

generalization to the *n*-person case is obvious, but for mutual knowledge in a population whose members aren't all acquainted with one another, I in effect proposed that *q* is mutual knowledge in *P* just in case everyone in *P* knows *q*, everyone in *P* knows that everyone in *P* knows *q*, and so on (Schiffer, 1972, §II.2). In his (1975) Lewis revised his account of common knowledge and said that a proposition is "*common* (or *mutual*) *knowledge* [in *P* just in case it] is known to everyone [in *P*], it is known to everyone [in *P*] that it is known to everyone [in *P*], and so on," but he added that the knowledge may be "merely potential: knowledge that would be available if one bothered to think hard enough." Conformity to a convention of truthfulness in *L* requires one not to utter any sentence of *L* unless it's true in *L*, but because "truthfulness-by-silence is truthfulness" (Lewis, 1969, pp. 165–166), it's not enough for a convention of truthfulness in *L* to prevail in *P* that members of *P* never utter sentences of *L*; in order for a convention of truthfulness to prevail in *P* members of *P* must regularly utter sentences of *L* and (for the most part) utter them only when they believe them to be true.

A striking feature of Lewis's account of the public-language relation is that, while it yields a definition of a sentence's having a certain meaning in a population, it says nothing at all about the meanings of words, nor does it say anything about a sentence's meaning being determined by the meanings of its constituent words and the semantic import of its syntactical construction. Lewis of course is well aware of the fact that:

> Not just any arbitrary infinite set of verbal expressions will do as the set of sentences of an interesting language. No language adequate to the purposes of its users can be finite; but any language usable by finite human beings must be the next best thing: finitely specifiable. It must have a finite grammar, so that all its sentences, with their interpretations, can be specified by reference to finitely many elementary constituents and finitely many operations for building larger constituents from smaller ones. (Lewis, 1969, p. 166)

He also of course recognizes that words as well as sentences have meanings for those who use them, and that the meaning of a word is determined, in so far as it is determined, by the way those for whom it has meaning use it. Why then doesn't Lewis define language in a way that requires a language to have a grammar, so that if a language *L* is used by a population *P*, then *L*'s words will have in *P* whatever meanings the grammar for *L* assigns them? Lewis explains that, while a grammar for *L* uniquely determines *L*, *L* doesn't uniquely determine any grammar: more than one grammar will determine *L*, grammars that may differ in the meanings they assign to the morphemes of *L* or even in what they recognize to be the morphemes of *L*. This presents Lewis with a problem:

> Given *P*, we select *L* by looking for a convention of truthfulness; but given *L*, how can we select [its grammar]? Conventions of truthfulness pertain to whole sentences and leave the interpretations of parts of sentences undetermined. Perhaps we should look for conventions of some other kind, but I cannot think what the content of such a convention might be. (Lewis, 1969, p. 198)

Well, perhaps the correct grammar for *L* is determined by something other than a convention. Lewis considers the Chomskian conjecture that the correct grammar for a language is the one that enters into the explanation of the linguistic competence of its users, but he rejects using that psycholinguistic hypothesis in defining the public-language relation; for

even if it's true, he says, it's a contingent truth, and thus can't be part of an *analysis* of '*L* is used by *P*,' "since the analysandum clearly could be true although the analysans was false" (Lewis, 1975, p. 178). Lewis is forced to conclude that he knows "of no promising way to make objective sense of the assertion that a grammar Γ is used by a population *P* whereas another grammar Γ′, which generates the same language as Γ, is not" (Lewis, 1975, p. 177).

As I have so far presented Lewis, one may be puzzled as to why he thinks grammars pose a problem for him. Why not identify languages with what he calls grammars, namely, finite specifications of the functions he now calls languages? Such a grammar may be conceived, at least initially, as a function that maps each expression of the language it determines onto its meaning in the language, and does so in a way that reveals how the meaning of every semantically complex expression is determined by its syntactical construction and the meanings the function assigns to the morphemes from which the complex expressions are constructed. Let L_G be such a Lewis-language-cum-grammar. In the functions Lewis defines as languages, each sequence of sounds in the function's domain is a sentence of the language. The sequences of sounds that constitute the domain of L_G will include words and other sub-sentential expressions along with sentences. If, for any ε and μ, $L_G(\varepsilon) = \mu$, then we may say that ε means* μ in L_G, and if L_G is a public language of population *P*, then ε means μ in *P*. For Lewis, a grammar determines the language of which it's a grammar, but a language doesn't determine its grammar. Since L_G is what Lewis calls a grammar, we would have a conception of language that does determine its grammar. Now suppose Lewis were to say that L_G is a public language of *P* just in case there prevails in *P* a convention of truthfulness in L_G, where, as before, there prevails in *P* a convention of truthfulness in L_G just in case it's common knowledge in *P* that its members try not to utter any *sentence* of L_G unless it's true in L_G, and so on. Then, provided the languages we speak are Lewis-languages, we can say what it is for a word *w* to mean μ in *P*: it means that just in case for some language L_G, L_G is a public language of *P*, *w* is a word in L_G, and $L_G(w) = \mu$. A Lewis-language maps only sentences onto meanings; let's call a finitely specifiable function L_G which maps expressions of every syntactic category onto a meaning a *Chomsky-language*. Every Chomsky-language determines a unique Lewis-language, but a Lewis-language determines no unique Chomsky-language. If there prevails in *P* a convention of truthfulness in a Lewis-language, then we have no way of saying that any grammar of the language is the grammar used in *P*. But if there prevails in *P* a convention of truthfulness in a Chomsky-language, then it follows that a particular grammar for a particular Lewis-language is used in *P*. We get that result simply by virtue of the fact that all it takes, according to Lewis, for there to be a convention of truthfulness in L_G in *P* is that it be common knowledge in *P* *that, for any sentence* σ *of* L_G *and proposition q, members of P try not to utter* σ *unless* $L_G(\sigma) = q$ *and q is true*, and so on.

Lewis didn't overlook this easy way of securing that a Chomsky-language is used in a population: he had already ruled it out by a stipulation he had made along the way about how the just-cited common knowledge condition was to be understood. Lewis says that a function *L* from finite sequences of sounds or marks – the "sentences" of *L* – into propositions is a public language of a population *P* only if it is common knowledge in *P* that members of *P* try never to utter a sentence of *L* unless it's true in *L*. But wouldn't such knowledge require ordinary speakers to know propositions of the form $L(\sigma) = q$, and wouldn't that require those ordinary speakers to know some set theory? And, just as bad, wouldn't knowing such propositions require them to have a finitely specifiable way of thinking about the function *L*, and wouldn't that be tantamount to knowing a grammar for *L*? Lewis has no doubt that ordinary

speakers have no such knowledge, and to avoid having it required by the common knowledge required for them to use a language, he first distinguishes between two ways in which the common knowledge in question might obtain in:

> It's common knowledge in *P in sensu composito* that there prevails in *P* a convention of truthfulness in *L* iff it's common knowledge in *P that for any σ, q such that L(σ) = q, a member of P won't utter σ unless she thinks q is true*.
>
> It's common knowledge in *P in sensu diviso* that there prevails in *P* a convention of truthfulness in *L* iff for any σ, *q* such that $L(\sigma) = q$, it's common knowledge in *P that a member of P won't utter σ unless she thinks q is true*.[2]

In other words, when the common knowledge is *in sensu composito*, what is known requires members of *P* to have a conception of *L* and enough knowledge of set theory to know propositions of the form $L(\sigma) = q$, whereas when the common knowledge is *in sensu diviso* members of *P* needn't have any way of thinking of *L* or any knowledge of set theory; they merely have to have the right expectations when sentences of *L* are uttered, so that if $L(\sigma) = q$ and a member of *P* hears another member of *P* utter σ, then she will expect him to believe *q*. "The common man," Lewis says, "need not have any concept of *L* in order to expect his fellows to be truthful … in *L*. He need only have suitable particular expectations about how they might act … in various situations" (Lewis, 1975, p. 180).

Now that we know why Lewis thinks he can say what it is for a sentence, but not for a word, to mean something in a population, and now that we have a better understanding of what, according to him, must be the case in order for a convention of truthfulness to prevail in a population, we are positioned to assess his claim that:

> For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff there prevails in *P* a convention of truthfulness in *L*.

There are, I believe, at least two problems with this account.[3] The first is that if there prevails in a population a convention of truthfulness in any language, then there will also prevail in that population conventions of truthfulness in infinitely many languages that are not public languages of the population. Suppose, for example, that English is the public language of *P*. Then members of *P* will regularly utter sentences of English, and when they do they will, for the most part, believe those sentences to be true. But only a finite number of the infinitely many English sentences will ever be uttered. Suppose ξ is an English sentence so convoluted and long that no one could reasonably expect it ever to be uttered. Then it will be obvious to any member of *P* who considers ξ that no member of *P* would utter it, and, *a fortiori*, for any proposition *r*, obvious that no member of *P* will utter ξ unless *r* is true. As noted earlier, with respect to the sentences of a population's language that will never be uttered, the truthfulness that obtains is truthfulness-by-silence. But now let Gobbledygook be a language that coincides with English with respect to every sentence that might be uttered but departs wildly from English thereafter: the sentences in Gobbledygook but not in English may be composed of words that aren't in English, or, if Gobbledygook and English have the same sentences, then the sentences that no one would ever utter have meanings in Gobbledygook that are entirely different from the meanings they have in English. Since it will be common knowledge (*in sensu diviso*) in *P* that members of *P* regularly utter sentences of Gobbledygook and that when they do they intend to be truthful in

Gobbledygook, and that they will be truthful-by-silence as regards the sentences of Gobbledygook that they know no one in *P* would ever utter, it follows by Lewis's definitions that there prevails in *P* a convention of truthfulness in Gobbledygook – a language that is not a public language of *P*.[4] And since infinitely many languages satisfy the description of Gobbledygook, it follows from Lewis's definitions that infinitely many languages constitute counter-examples to his definition of the public-language relation.

I mentioned this problem to Lewis in 1968; he acknowledged that it showed that his definition failed to provide a sufficient condition, but he couldn't at the time find a revision that avoided the problem. That may seem surprising. The counter-example works only if one takes '$\neg A$ unless B' to be equivalent to the material condition '$A \rightarrow B$,' which one can know to be true merely by knowing that A is false. But if Lewis were to say instead that the common knowledge is of the *counterfactual* proposition *that if a member of P* ***were*** *to utter ξ, then she* ***would*** *think that q was true*, then Gobbledygook wouldn't be a counter-example to the definition, for if, for any σ, $G(\sigma) \neq E(\sigma)$, then no member of *P* would expect a member of *P* to mean $G(\sigma)$ if she were to utter σ, for they wouldn't know to associate that proposition with σ. The problem with this thought, however, is that there are also infinitely many English sentences that members of *P* wouldn't be able to understand – sentences like 'Buffalo buffalo buffalo buffalo buffalo buffalo buffalo buffalo buffalo,'[5] in addition to the infinitely many English sentences whose length or convoluted structure make them impossible to process. In "Language and languages" (Lewis, 1975), Lewis proposed a certain way of avoiding the problem, but then in "Meaning without use: reply to Hawthorne" (Lewis, 1992), he recognized that that solution to what he now called the *meaning-without-use problem* didn't work, and proposed a third account of the public-language relation that, as we'll see in §4, may also be problematic.

The second objection to *Convention*'s account of the public-language relation also shows that that account fails to provide a sufficient condition for a language to be a public language of a population. The problem is that if it were a sufficient condition for *L*'s being a public language of *P* that there prevailed in *P* a convention of truthfulness in *L*, then virtually *every* convention would count as a convention of truthfulness in a language, and this by virtue of the fact that every (or virtually every) convention requires certain actions to be performed when certain conditions obtain. For example, suppose that in monastery *M* there is a convention to recite prayer *A* on Monday, prayer *B* on Wednesday, and prayer *C* on Friday, and not to recite those prayers on any other days. Now let $L^{\dagger}$ be that function such that:

$$\forall x, y \left[L^{\dagger}(x) = y \text{ iff } (1)\ x = \text{reciting } A\ \&\ y = \text{the proposition that it's Monday or } (2)\ x = \text{reciting } B\ \&\ y = \text{the proposition that it's Wednesday or } (3)\ x = \text{reciting } C\ \&\ y = \text{the proposition that it's Friday}\right]$$

Then $L^{\dagger}$ is a Lewis-language and there prevails in *M* a convention of truthfulness in $L^{\dagger}$, and Lewis is committed to saying that $L^{\dagger}$ is used in *M* as a public language of communication. But since the members of *M* never use $L^{\dagger}$ to communicate anything (at least to one another), it's clearly not used as a public language of communication (or as any other kind of language). I mentioned this problem, too, to Lewis in 1968; he agreed that it was a counter-example to his definition and changed that definition to:

> For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff there prevails in *P* a convention of truthfulness in *L, sustained by an interest in communication*,

which is the definition that appeared in *Convention* when the book was published. The revision, however, doesn't do the job: prayers *A*, *B*, and *C* might all be for greater powers of communication; yet that still wouldn't make $L^{\dagger}$ a language the monastery uses as a public language of communication. A convention must forge a considerably tighter connection to communication if it's to succeed in defining the public-language relation; it would at least have to secure that sentences of the language it concerns are uttered in order for speakers to mean what the sentences mean* in the language. This is our segue to the Gricean program of intention-based semantics.

2 Intention-Based Semantics

This is a program for reducing all questions about the intentionality of speech acts and linguistic expressions to questions about the intentionality of thought. It takes as foundational in the theory of meaning a certain notion of *speaker-meaning* and seeks to define it, without recourse to any semantic notions, in terms of acting with certain audience-directed intentions. Then it seeks to define other agent-semantic notions – most notably, speaker-reference (the notion of a *speaker*'s referring to a thing, as contrasted with an *expression*'s referring to it) and illocutionary acts[6] – in terms of its defined notion of speaker-meaning.[7] With that done, Gricean IBS then sets out to define the semantic features of linguistic expressions wholly in terms of its defined notion of speaker-meaning, together with ancillary notions, such as that of *convention*, which are themselves explicable wholly in terms of non-semantic propositional attitudes. Since expression-meaning is defined in terms of speaker-meaning and convention, and speaker-meaning and convention are defined in terms of non-semantic propositional attitudes, it's supposed to follow that expression-meaning is also defined in terms of non-semantic propositional attitudes.

There is disagreement among Griceans as to how exactly the definition of speaker-meaning should go, but there may be little reason to care about how to define a notion of speaker-meaning if that notion can't be used to define the semantic features of linguistic expressions, and in this regard a reader of Grice's 1957 article "Meaning" has a right to be puzzled. Grice spends nearly all of that article building up to his famous proposal that *S* meant something in "uttering" *x* iff *S* "intended the utterance of *x* to produce some effect in an audience by means of the recognition of this intention,"[8] and it is only at the very end of the article that we get anything about expression-meaning. But then all we get are two equivalences baldly presented without any elaboration. The first is that:

> '*x* meant something' is (roughly) equivalent to 'Somebody meant something [in uttering] *x*.'

For example, if Gretel rolls her eyes to communicate to Hansel that the speaker is a pretentious bore, then, by the definition, Gretel's eye rolling meant something. The second generalization is that:

> '*x* means (timeless) that so-and-so' might as a first shot be equated with some statement or disjunction of statements about what "people" (vague) intend (with qualifications about "recognition") to effect by *x*.

There are four things one is apt to find puzzling. First, Grice has made no effort to show how these definitions are motivated by his definition of speaker-meaning. Second, the

equivalence offered for '*x* meant something' seems not to be an analysis of any obvious pre-theoretic notion, but is more in the nature of a stipulation whose theoretical purpose hasn't been revealed. Third, there is no mention of word-meaning. And fourth, the second equivalence, which is intended to cover indicative sentence meaning, appears to ignore the fact that every natural language has infinitely many sentences that will never be, or even could be, uttered. So why is Grice's article so famous and thought by many to be of such importance? Does the importance of Grice's article reside wholly in his suggested account of speaker-meaning, never mind any relevance that account might have for an account of expression-meaning?

No; those who perceived Grice's article to be important did so because they took themselves to discern in his account of speaker-meaning an *invisible hand* that guides them from the intentions that define speaker-meaning to an account of expression-meaning in terms of those intentions. Any Gricean account of speaker-meaning could be used to make the invisible hand visible, but I will use the account of assertoric speaker-meaning implicit in Grice's 1957 article. According to that account:

> For any person *S*, proposition *p*, and utterance *x*, *S* meant *p* in uttering *x* iff, for some person *A*, *S* uttered *x* intending
>
> (1) *A* to believe *p*;
> (2) *A* to recognize that *S* uttered *x* intending (1);
> (3) *A*'s recognition of that intention to function as part of *A*'s reason for believing *p*.

Quite apart from the question of whether this is in any way "correct," its proper understanding requires seeing the answers Grice assumed to two questions raised by the definition. One question was how *A*'s recognition of *S*'s intention to get *A* to believe *p* was supposed to function as part of *A*'s reason for believing *p*. The intended answer was that *A* would infer from the fact that *S* uttered *x* intending *A* to believe *p* that *S* believed she knew *p*, and then, taking the fact that *S* believed she knew *p* to be very good evidence that *p* was true, infer *p* from that fact.[9] The other question raised by the above displayed definition was how *A* was to recognize that *S* uttered *x* intending *A* to believe *p*. Understanding how that recognition was supposed to work is essential to understanding the invisible hand that was supposed to guide one from the account of speaker-meaning to an account of expression-meaning. It was supposed to work like this: when *S* utters *x* in order to mean *p*, there is some feature φ such that *S* intends *x* to have φ and intends *A* to recognize that *x* has φ and to infer in part therefrom that *S* uttered intending *A* to believe *p*. Making this explicit yields the following slightly tweaked version of Grice's account of assertoric speaker-meaning:

> For any person *S*, proposition *p*, and utterance *x*, *S* meant *p* in uttering *x* iff for some feature φ and person *A*, *S* uttered *x* intending
>
> (1) *x* to have φ;
> (2) *A* to recognize that *x* has φ;
> (3) *A*'s recognition that *x* has φ to function as at least part of *A*'s reason for believing that *S* uttered *x* intending:
> (4) *A* to believe *p*;
> (5) *A*'s recognition of *S*'s intending *A* to believe *p* to function as at least part of *A*'s reason for believing *p*.

Let's call the value of 'φ' in an act of speaker-meaning its *inference-base feature* (its *IB-feature*, for short). Acts of speaker-meaning are typically performed by uttering sentences of a language common to the speaker and her audience, and the IB-features of those sentences are their meanings. If you utter 'It's snowing' to communicate that it's snowing, the IB-feature of the sentence 'It's snowing' on which you rely is the meaning of that sentence in English. The reason IB-features are typically meaning properties is that meaning properties are optimal IB-features: if you want to tell your child that it's snowing, you would do much better to utter 'It's snowing' than to attempt to communicate that it's snowing by impersonating a snow flake or uttering 'The flamingoes are flying south early this year.' At the same time, a *sine qua non* of a Gricean account of speaker-meaning is that the only intentional notions mentioned on its right-hand side are ordinary propositional-attitude notions, and, consequently, it's not a necessary condition for a person's meaning a proposition that what she utters have a semantic property as its IB-feature. For example, during a lecture one might communicate to one's friend that one is bored by closing one's eyes and pretending to snore. The invisible hand in the displayed account of speaker-meaning is that in any population whose members have frequent need to communicate with one another, the fact that they frequently communicate with one another will result in their utterances having IB-features that are both optimal and specifiable wholly in terms of what they know their speaker-meaning practices to be. The invisible-hand idea is that, since the notion of speaker-meaning utilized in these optimal IB-features is defined in wholly non-semantic terms, these features will also be intrinsically specifiable in wholly non-semantic terms. The Gricean invisible-hand strategy is completed by *identifying* meaning properties with those non-semantically specifiable optimal IB-features, thereby explaining the optimality of meaning properties as IB-features.

The Gricean account of assertoric speaker-meaning requires one further tweak before we have an account of assertoric speaker-meaning that best reveals the Gricean strategy for explaining expression-meaning in terms of the conditions that define speaker-meaning. Earlier (see §1) I mentioned the notion of *common knowledge* Lewis introduced in *Convention* and the similar notion of *mutual knowledge* that I introduced in *Meaning*. The further tweak is that mutual (or common) knowledge must be added to the Gricean mix to yield this account of speaker-meaning:

> [SM] For any person *S*, proposition *p*, and utterance *x*, *S* meant *p* in uttering *x* iff for some feature φ and person *A*, *S* uttered *x* intending it to be mutual knowledge between *S* and *A* that *x* has φ and, at least partly on that basis, mutual knowledge that *S* uttered *x* intending *A* to believe *p* and intending their mutual knowledge that *S* uttered *x* intending *A* to believe *p* to be at least part of *A*'s reason for believing *p*.

Mutual knowledge was originally introduced to repair the failure of Grice's original conditions to provide a set of jointly sufficient conditions for speaker-meaning (see, e.g., Schiffer, 1972, II.1 and 2); I invoke it now because of the way it will be needed in the Gricean attempt to define expression-meaning in terms of speaker-meaning. There are better and worse ways to understand mutual knowledge, and I might not have opted for the best way in my book. The essential job mutual knowledge needs to perform is to capture the sense in which acts of communication require the defining features of speaker-meaning to be "out in the open" between speaker and hearer. In *Meaning* I offered a set of finite conditions for generating mutual knowledge, and I now think that I would have done best simply to have identified mutual knowledge with a version of those base conditions. In any case, for present

purposes I'll continue to use 'mutual knowledge' and its cognates as dummy expressions for whatever turns out to be the best accommodation of the requisite out-in-the-openness.

The Gricean takes his invisible-hand strategy to be most clearly and paradigmatically exhibited in his account of simple signals,[10] and a little thought experiment will show how that is supposed to work. There is a weekly seminar regularly attended by the same people. Their practice is to raise their hands if they want to be called on, but, while they would benefit from one, they have no simple way of indicating that what they want to contribute is a follow-up question. During one session, a visitor from the University of Latvia, Zuzka, raises her hand during a lively discussion and moves her index finger rapidly up and down. It's clear to all that Zuzka intends to communicate something by this gesture, but at first no one can figure out what it is. After several minutes it somehow transpires that in Latvian universities that sort of finger movement means that one has a follow-up question. Now suppose that during another exchange the following week one of the attendees, Harvey, raises his hand and moves his index finger rapidly up and down. In this case, everyone in the class will know straightway that Harvey means that he has a follow-up question. What explains this dramatic difference? Why was it that no one knew what Zuzka meant in moving her finger in way **Ω**, whereas one week later the very same people effortlessly and immediately knew that in moving his finger in way **Ω** Harvey meant that he had a follow-up question? The answer, of course, is that at the time Zuzka performed the finger movement it had no feature that was an effective IB-feature in the seminar for meaning that one had a follow-up question, but after that the movement had a few features it didn't previously have, and these features separately and together constituted quite an effective IB-feature for meaning that one had a follow-up question. One of these features was that of being mutually known to be such that it was performed by Zuzka in her attempt to communicate to the seminar that she had a follow-up question; another was that of being mutually known to be the standard way in Latvian universities to communicate *that one had a follow-up question* (*Q*, for short). Since it benefited the seminar to have a simple way to indicate that one had a follow-up question, now that no one doubts what one would mean by moving one's index finger in way **Ω**, it's apt to catch on, so that it soon becomes mutual knowledge in the seminar that there is a practice of meaning *Q* by moving one's index finger in way **Ω**, and that mutual knowledge makes the gesture an optimal IB-feature as regards meaning *Q* in the seminar.

You may recognize that we have entered the territory of the kind of self-perpetuating regularities that David Lewis showed to be conventions, and at this point it seems correct to say that the gesture means *Q* in the seminar. Now the gesture **Ω** doesn't mean a proposition, since each person who makes the gesture will mean that *she* has a follow-up question. But to keep what is essential to the lines we are exploring from being hidden in complexities and qualifications we are ignoring non-assertoric speech acts, indexicality, and ambiguity, and pretending that sentence-size meanings are propositions. Relative to all that, the Zuzka example suggests an account of simple-signal meaning that entails the following:

[SIMP] For any *x*, proposition *q*, and population *P*, *x* is a simple signal that means *q* in *P* iff it's mutual knowledge in *P* that there is a practice in *P* of meaning *q* by uttering *x* and intending that mutual knowledge to function as *x*'s IB-feature when a member of *P* means *q* by uttering *x*.

Practices that satisfy SIMP are conventional practices (see Schiffer (1972, V.3), but it may not be easy to describe them as conventions. At least SIMP doesn't require there to be a

convention in P to mean q by uttering x, which is good, since x can mean q in P even though members of P have ways other than uttering x to mean q, nor does it require there to be a convention in P not to utter x unless one thereby means q, which is also a good thing, for even if x means q in P members of P might communicate other propositions by, for example, using x metaphorically. In any case, whether or not the practice entailed by SIMP can be described as a convention, as opposed to a conventional practice, that wouldn't give us reason to suppose SIMP fails to provide either a necessary or a sufficient condition for simple-signal meaning.

The Gricean sees an even broader application for the kind of IB-feature employed in SIMP. Recall that for the Gricean speaker-reference is to be defined in terms of speaker-meaning, and this will be so in a way that suggests that the name-of relation is best captured by something along the lines of:

> [*N*] *n* is a name of *y* in *P* iff it's mutual knowledge in *P* that there is a practice in *P* of speakers' referring to *y* with *n* & intending that mutual knowledge to be *n*'s IB-feature.[11]

In other words, when you say to me 'Saul Kripke is giving a talk today' you intend the feature of 'Saul Kripke' that enables me to know that you are referring to Saul Kripke to be that it's mutual knowledge in a population to which we both belong that there is a practice of referring to Kripke with 'Saul Kripke,' which for the Gricean is roughly equivalent to saying that what enables me to know that you are referring to Kripke with 'Saul Kripke' is that it's mutual knowledge between us that 'Saul Kripke' is the name of Saul Kripke in a population to which we both belong.

Of course, the Gricean can't hope to account for the meanings of natural language sentences in a similar way, since a sentence has its meaning even if no one has ever uttered it. How, then, might the Gricean invisible-hand strategy apply to natural languages? The Gricean expects that if σ means q in the language L of a population P then, while there needn't be any practice in P of uttering σ, there will prevail in P a set of practices pertaining to L such that one utters σ in conformity with those practices only if one means q in uttering σ. The question is, what might those practices be? Unfortunately for the Gricean, we are already positioned to see that his goal of defining expression-meaning in terms of his defined notion of speaker-meaning is an impossible goal for him to achieve. The invisible-hand strategy fails to take one to any good place. We can appreciate this in the following way.

SIMP, the lately displayed Gricean definition of simple-signal meaning, is apt to seem plausible. The only semantic notion used in SIMP is the notion of speaker-meaning, and SM, which I'm taking to represent the Gricean account of speaker-meaning, has been defined in terms of non-semantic propositional attitudes. It may therefore seem that, if both SIMP and SM are correct, then the Gricean has succeeded in defining simple-signal meaning in wholly non-semantic terms, and thereby at least to have shown that the intentionality of simple signals derives from the intentionality of the propositional attitudes in terms of which it's defined. *Not so!* Even if SIMP and SM are correct, that would not entail that simple-signal meaning had been indirectly defined in terms that included those that define speaker-meaning. The point is familiar and simple. Suppose that

$$A =_{\text{def}} \ldots X \ldots Y \ldots$$

and that

$$B =_{\text{def}} \ldots \boldsymbol{A} \ldots Z \ldots$$

then

$$B =_{\text{def}} \ldots X \ldots Y \ldots Z \ldots$$

does *not* follow *if the context of 'A' in 'B =$_{def}$ … **A** … Z …' is intentional*. Specifications of knowledge are intentional contexts *par excellence*, and *speaker-meaning is mentioned in SIMP only in specifying the mutual knowledge required for simple-signal meaning*. Consequently, we have no reason to expect that if SIMP is correct, then SIMP$_G$, the result of replacing every speaker-meaning expression in SIMP with what SM provides as its definitional expansion, will also be correct. That it would be wrong to expect SIMP$_G$ to be true if SIMP is true would seem to be confirmed by a side-by-side comparison of the two definitions:

SIMP	**SIMP$_G$**
For any *x*, proposition *q*, and population *P*, *x* is a simple signal that means *q* in *P* iff it's mutual knowledge in *P* that there is a practice in *P* of meaning *q* by uttering *x* and intending that mutual knowledge to function as *x*'s IB-feature when a member of *P* means *q* by uttering x.	For any *x*, proposition *q*, and population *P*, *x* is a simple signal that means *q* in *P* iff it's mutual knowledge in *P* that there is a practice in *P* whereby a member *S* of *P* utters *x* in order that there be something *y* that *S* utters such that, for some feature φ and member *A* of *P*, *S* utters *y* intending it to be mutual knowledge between *A* and *S* that *y* has φ and, at least partly on that basis, mutual knowledge that *S* uttered *y* intending *A* to believe *q* and further intending their mutual knowledge that *S* uttered *y* intending *A* to believe *q* to be at least part of *A*'s reason for believing *q*, and in such utterances of *x S* intends the inference-base feature φ of *x* to be the fact that it's mutual knowledge in *P* that there is a practice in *P* whereby a member *S* of *P* utters *x* in order that there be something *y* that *S* utters such that, for some feature φ and member *A* of *P*, *S* utters *y* intending it to be mutual knowledge between *A* and *S* that *y* has φ and, at least partly on that basis, mutual knowledge that *S* uttered *y* intending *A* to believe *q* and further intending their mutual knowledge that *S* uttered *y* intending *A* to believe *q*.

SIMP is apt to seem plausible, SIMP$_G$ quite implausible, which isn't surprising, given that we have no reason at all to expect the two definitions to have the same truth-value even if SIMP and SM are both true.[12]

The Gricean invisible-hand strategy for defining the semantic properties of expressions in non-semantic terms never gets off the ground: the Gricean simply has no way of defining expression-meaning in terms of the *intentions* that by his lights are constitutive of

speaker-meaning. But that doesn't preclude defining expression-meaning in terms of the *concept* of speaker-meaning, and that suggests the possibility of a convention-based way of reducing the semantic to the psychological that isn't strictly Gricean *intention*-based semantics but might be considered a cousin of it and, if correct, would show how the intentionality of language derives from the intentionality of thought. The next section explores this idea.

3 An Almost-Gricean Semantics

For all we yet know, SIMP, which defines simple-signal meaning in terms of a certain kind of conventional speaker-meaning practice, remains plausible. All that was shown was that the attempt to replace the speaker-meaning expressions in SIMP with their Gricean expansions results in an implausible account of simple-signal meaning, even if the Gricean account of speaker-meaning is correct. One familiar diagnosis of this substitutivity failure which derives from Frege is that when an expression occurs in an intentional context, as when it's used to specify what someone knows, the expression doesn't denote the object, property, or relation it denotes when it occurs in extensional contexts, but denotes instead a *concept* of what it denotes in those extensional contexts. The idea, in other words, is that the occurrence of 'means' in the sentence 'In making that gesture Zuzka meant that she had a follow-up question' denotes the speaker-meaning relation, but its occurrence in an intentional context, such as the context created by 'knows' in 'Sid knows that in making that gesture Zuzka meant that she had a follow-up question,' denotes a *concept* of that relation and not the relation itself. Suppose that, or something enough like it, is right. Then we would have defined simple-signal meaning not, as the Gricean would have it, in terms of the intentions that define speaker-meaning, but rather in terms of what, for all we yet knew, was a primitive concept of speaker-meaning (see Schiffer, 1987, §9.3). But *if* (i) that concept of speaker-meaning could be identified with a psychological construct that was itself definable in non-intentional terms – say, in terms of the inferential and causal roles of neural expressions in mentalese, (ii) the speaker-meaning relation it denoted was definable without recourse to any intentionality other than the intentionality of non-semantic propositional attitudes, and (iii) we can move beyond SIMP to give an account of the semantic properties of all expressions in terms of conventional practices whose specification involved that concept of speaker-meaning, *then* we could achieve something like a reduction of the semantic to the psychological. And if, further, the speaker-meaning relation denoted by the concept of speaker-meaning was definable *à la* Grice and meaning properties could still be conceived as IB-properties, then – what the hell – we would have a theory that was close enough to Gricean semantics to be called an almost-Gricean semantics.

To see whether that can be done, a good place to begin would be to assume that SIMP, the account of simple-signal meaning, was correct and try to see whether speaker-meaning conventions (or conventional practices) might fare better than conventions of truthfulness in a definition of the public-language relation. We are encouraged to think that might be so by the problem with which we left Lewis's proposal that

> For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff there prevails in *P* a convention of truthfulness in *L*, sustained by an interest in communication.

This was the problem that virtually every convention – for example, a convention to drive on the right or to wear casual clothes to work on Fridays – was for Lewis a convention of truthfulness in a "language" that was in no sense used as a medium of communication,

not even when the convention was somehow or other "sustained by an interest in communication." The encouragement this problem gave to a Gricean approach was that it suggested that in order for *L* to count as a public language of *P* the convention (or conventional practice) governing the use of *L* should be a convention conformity to which requires uttering sentences of *L* in order to mean what those sentences mean* in *L*. In pursuing this line of thought, however, it's important to keep in mind that, while a correct definition of the public-language relation would *eo ipso* be an account of what it was for a sentence to have meaning in a population, that would be no stopping point even for almost-Griceans unless the account of the public-language relation, unlike those Lewis had on offer, somehow or other managed also to define what it was for an expression of *any* syntactic kind – words and phrases, as well as sentences – to have meaning in a population.

The most obvious Gricean counterpart to Lewis's notion of a convention of truthfulness would be a convention of meaning, where that is taken to mean that there prevails in *P* a convention not to utter any sentence σ of *L* unless, for some *q*, $L(\sigma) = q$ and in uttering σ one means *q*, and where that in turn entails mutual knowledge which, when understood *in sensu diviso*,[13] comes to:

(A) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff for any σ, *q*, if $L(\sigma) = q$, then it's mutual knowledge in *P*
  (1) that a member of *P* won't produce an unembedded utterance of σ unless she means *q* thereby;[14]
  (2) that if a member of *P* means *q* in uttering σ, then she intends σ's IB-feature to be the fact that it's mutual knowledge in *P* that (1).

(A) has more than a few problems, the most worrisome of which are, first, that it suffers from the same meaning-without-use problem that (A)'s counterpart in Lewis was seen to suffer from and, second, that it doesn't explain what it is for a word or other sub-sentential expression to have meaning in a population. Brian Loar did as much as any Gricean to understand expression-meaning in Gricean terms, and he offered an account of the public-language relation that both took into account the problems that arose for Lewis's account and provided an account of what it is for an expression of any syntactic kind to have meaning in a population (Loar, 1976). Taking our cue from Loar, and ignoring some fixable problems, we need to take three more steps to achieve what by Loar's lights would be a correct account of the public-language relation (relative to our ongoing simplifications about ambiguity, vagueness, indexicality, non-indicative moods, and every sentence having a proposition as its meaning).

The first step is to counterfactualize the mutual knowledge in (A), thereby getting:

(B) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff for any σ, *q*, if $L(\sigma) = q$, then it's mutual knowledge in *P*
  (1) that if a member of *P* were to produce an unembedded utterance of σ she would mean *q* thereby;
  (2) that if a member of *P* means *q* in uttering σ, then she intends σ's IB-feature to be the fact that it's mutual knowledge in *P* that (1).

The problem with (B), as Loar noticed, is one already encountered in our discussion of Lewis: it requires speakers of a language to know the meaning of every sentence of their

language, whereas we already know that there are infinitely many sentences of every spoken language that are too long or convoluted for ordinary speakers to understand. The fix Loar proposes, which takes us to the second step, restricts the mutual knowledge to those sentences of their language members of a population can understand:

(C) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff *there is a large enough restriction L′ of L* such that, for any σ, *q*, if $L'(\sigma) = q$, then it's mutual knowledge in *P*
  (1) that if a member of *P* were to produce an unembedded utterance of σ she would mean *q* thereby;
  (2) that if a member of *P* means *q* in uttering σ, then she intends σ's IB-feature to be the fact that it's mutual knowledge in *P* that (1).

Loar knew that (C) was no stopping point, for it has the same meaning-without-use problem that Lewis's definition of the public-language relation has: infinitely many languages that aren't public languages of any population will each entail the restriction *L′* before differing wildly from one another beyond what they have in common; and, of course, it doesn't explain what constitutes a word's having meaning in a population.

The languages we are trying to capture are finitely specifiable. A finite specification of a language is a grammar. In *Convention*, Lewis argued that, if the infinite language *L* is the language of *P*, then *L* must be finitely specifiable, but there will be no fact of the matter as to which finite specification of *L* is the correct specification of *L*. Lewis reverses himself in (1992). There he comes to the realization that in order to determine which language is a population's public language it must be possible to determine which grammar generates that language, but he now thinks he knows how this can be done:

> True, there are many grammars. But they are not on equal terms. Some are "straight" grammars; for example, any grammar that any linguist would actually propose. Others are "bent," or "gruesome," grammars; for example, what you get by starting with a straight grammar for English and adding one extra rule, which states that every expression with more than forty occurrences of the word 'cabbage' is a sentence meaning that God is great. We have no difficulty in telling the difference …. We can reasonably hope that all straight grammars that agree on the used fragment will agree everywhere. (Lewis, 1992, p. 110)

The application of this solution to (C) yields:

(D) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff there is a large enough restriction *L′* of *L* and a grammar Γ such that (a) Γ is the "straightest" grammar that determines *L′*, (b) Γ determines *L*, and (c) for any σ, *q*, if $L'(\sigma) = q$, then it's mutual knowledge in *P*
  (1) that if a member of *P* were to produce an unembedded utterance of σ she would mean *q* thereby;
  (2) that if a member of *P* means *q* in uttering σ, then she intends σ's IB-feature to be the fact that it's mutual knowledge in *P* that (1).

A striking feature of this solution to the meaning-without-use problem is that it doesn't require the grammar that determines a person's language to play any role in the psycholinguistic explanation of the information processing that underlies, and thus accounts for, the

person's ability to understand utterances in that language. The striking fact invites an objection to the solution, which I raised when Lewis and I discussed a version of his (1992) in 1990 – namely, that if we learned that the internally represented grammar implicated in the explanation of a person's ability to understand novel sentences was in fact a bent grammar that determined a language $L^{\#}$, then we should want to say that $L^{\#}$ was her language, even if the only straight grammar that fit the used fragment determined a different language. Lewis was unmoved:

> Maybe there is a grammar somehow written into the brain. And conceivably it is a bent grammar, so that the language it generates differs, somewhere outside the used fragment, from the language we get by straight extrapolation. Schiffer has asked: does straight extrapolation give the right answers even then? I think so. If not, then whenever we resort to extrapolation to answer questions of syntax and semantics, we are engaged in risky speculation about the secret workings of the brain. That seems wrong. (Lewis, 1992, p. 110, n. 6)

Risky speculation about the secret workings of the brain is certainly to be avoided, but, as we'll presently see, it's not entailed by the hypothesis that the criterion for being a population's language is that it's determined by the grammar that explains how members of the population are able to understand the utterances they hear. For suppose that the infinite language L is in fact the public language of population P. Let L' be the fragment of L that has been or might be produced in P. We may assume that each member of P has a compositional understanding of L', one that relies on the only straight grammar to fit L' and the only grammar that both fits L' and determines L. We should certainly insist that L is the language of P. Now it's possible for there to be another population $P^{\star}$ such that L' is also the fragment of their language that has been or might be produced in $P^{\star}$, but with this difference: each sentence of L' is for them a *non-composite* utterance type. In other words, each sentence of L' is for them as a simple signal, such as a fire alarm, is for us: it has propositional meaning, but its meaning is not in any way determined by semantic features of its parts and structure. The "language" L' is simply a large finite set of simple signals; the members of $P^{\star}$ have prodigious memories, and they have learned the sentences of L' the only way they could learn them – namely, one by one. They know what their sentences mean because they have learned what each one means as a single fact, and they have no way of computing what a sentence means on the basis of their knowledge of its syntax and the meanings of its parts, as, from their perspective, a sentence no more has a syntax and semantically relevant parts than a fire alarm has for us. Consequently, members of $P^{\star}$ have no way of understanding a sentence that belongs to L but not to L'; they have no way, in fact, of determining the meaning of any novel sentence. Clearly, the infinite language L is *not* used in $P^{\star}$. Yet every straight grammar that generates L', the language that *is* used by them, is a grammar of L. There are also counter-examples to the straight-grammar solution involving infinite languages. Suppose, for example, that the members of a secret society make up a language whose syntax and semantics they actually write down and formally adopt under a Mafia-like oath, and that, for some reason, perverse or not, the grammar they adopt is by their explicit design a bent grammar. Let's also suppose that an internal representation of the bent grammar plays an essential role in their processing of utterances in their invented language. I would think that the language of the society is the one their bent grammar describes, whether or not

there is a straight grammar that fits the sentences the members of the society actually produce or are likely to produce.

Brian Loar has suggested that appealing to a grammar that is somehow written into the brain is exactly what must be done both to solve the meaning-without-use problem and to achieve an account not just of a sentence's meaning in a population but of every expression's meaning in a population. He first stipulates that "*L* is grounded in *P*, with regard to its restriction *L′*, just in case those correlations of sentential features and meaning contributions which figure in the correct psychological explanation of the continuing mastery of *L′* (i.e. effective *L*) by members of *P* will generate, when extended, the full language *L*, including its incomprehensibly complex sentences" (Loar, 1976, pp. 159–160). This would yield:

(E) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff there is a large enough restriction *L′* of *L* such that, for any σ, *q*, if *L′*(σ) = *q*, then (a) it's mutual knowledge in *P*
  (1) that if a member of *P* were to produce an unembedded utterance of σ she would mean *q* thereby;
  (2) that if a member of *P* means *q* in uttering σ, then she intends σ's IB-feature to be the fact that it's mutual knowledge in *P* that (1)
  and (b) *L* is grounded in *P* with regard to *L′*.

But what about Lewis's objection that if the grammar that determines the used language is the one written into the brain, then "whenever we resort to extrapolation to answer questions of syntax and semantics, we are engaged in risky speculation about the secret workings of the brain"? Well, suppose I believe that the correct grammar for my language is whatever grammar is written into the brains of those of us who use the language. It would be no small feat to construct a grammar that merely fits the fragment of my language that has been or is ever likely to be used, but suppose that, after considerable labor, experts come up with a grammar Γ that gets the fragment right. Γ will determine not just the fragment in question but an entire infinite language: it assigns a syntactic, semantic, and phonological interpretation to each of infinitely many sentences, all but a small minority of which have no chance of ever being used by me or anyone else. It seems a very good bet that Γ is a straight grammar, probably the only straight grammar anyone can come up with that fits the used fragment. So now I must decide: Is Γ the grammar implicated in my language processing, or is that grammar one that coincides with Γ on the used fragment but differs from it in that it interprets every sentence containing 40 or more occurrences of 'cabbage' as meaning that God is great? I will of course go with the straight grammar Γ, but not on the basis of a risky speculation. My going with the straight grammar Γ is a risky speculation about the secret working of the brain only if the scientist who infers a straight theory from the evidence that underdetermines it, rather than one of the infinitely many bent theories that fits the same evidence, isn't making a risky speculation about the infinitely many possible unexamined cases her theory must cover.

Nevertheless, there is a serious problem with (E) or any other almost-Gricean account of the public-language relation which avoids the meaning-without-use problem by appeal to the grammar implicated in the explanation of our language processing. The problem is that if, like Loar, one invokes groundedness in one's account of the public-language relation – that is to say, into one's account of what determines an expression to have meaning in a population – then one is invoking something that in

effect makes the rest of one's account superfluous. For if one is invoking groundedness, then one needs nothing more than:

(F) For any Lewis-language *L* and population *P*, *L* is a public language of *P* iff (1) members of *P* regularly communicate with one another by uttering sentences of *L* and meaning thereby what those sentences mean* in *L*, and (2) for some grammar Γ of *L*, their ability to know what members of *P* mean in uttering those sentences is grounded in Γ.[15]

(F) is neither Gricean nor convention-based; it uses the notion of speaker-meaning, but relies on no *particular* account of it, Gricean or otherwise. I'm not suggesting that (F) is correct. My point is merely that it would appear that in order for an almost-Gricean convention-based theorist to have an account of the public-language relation (= an account of expression-meaning) she will in effect need Loar's notion of groundedness, which entails the right-hand side of (F), and that by anyone's lights should already be a necessary and sufficient condition for a Lewis-language's being a public language of a population.[16]

4 What Endures?

In this chapter I have looked critically at two programs that should dominate any curriculum bearing the label "Intention and Convention in the Theory of Meaning": David Lewis's convention-based account of sentence meaning and the more comprehensive Grice-inspired program of intention-based semantics. As for Lewis, one may object to his account of convention (see, e.g., Gilbert, 1981) and one may deny that the languages we speak can be represented as Lewis-languages (Lewis, 1975, p. 180), but we have seen that even if Lewis's account of convention is correct and the languages we speak can be represented as Lewis-languages, no convention or conventional practice – whether of truthfulness or of speaker-meaning – can explicate what it is for an expression to have meaning for a person or population of persons.

The more ambitious IBS also fails in its attempt to explain expression-meaning. First, it simply doesn't follow that if one can define expression-meaning in terms of speaker-meaning and speaker-meaning in terms of intentions that one can thereby define expression-meaning in terms of intentions. That would follow only if mention of speaker-meaning in the analysis of expression-meaning occurred in a non-intentional context, whereas its occurrence is in a specification of the mutual knowledge the account requires, a paradigm intentional context. Second, as we saw in our discussion of almost-Gricean IBS, an expression's meaning in a population can't be defined in terms of the speaker-meaning practices that prevail in the population. What of the possibility of a Gricean account of speaker-meaning? True, the limitations we now see on the work it could do considerably diminish the interest in having such an account, but nothing has been said so far in this chapter to show that no such account can be correct. And while such an account couldn't be used to explicate expression-meaning, it's not unreasonable to suppose it would have relevance to speech-act theory and to issues about the semantics/pragmatics interface, especially since Grice is nowadays best known for his theory of implicature, and for Grice (nearly enough)

S implicates p in uttering σ iff in uttering σ *S* means but doesn't say *p*,

where in Grice's quasi-technical sense of 'say,'

S said p in uttering *σ* iff *p* "fits" the meaning of σ and *S* meant *p* in uttering σ,

where "fitting" is exemplified in, say, the way the proposition *that Betty is ready to take the exam* fits the meaning of 'She is ready.' And there remains the unsettled question of whether semantic intentionality can be reduced to the intentionality of mental states.

There are reasons to doubt that a Gricean account of speaker-meaning can be achieved. First, it's not even clear what an account of "speaker-meaning" is supposed to be an account of when it's not constrained by the need to be the speech-act notion needed to define expression-meaning.[17] Grice's "Meaning" (and, ahem, my own *Meaning*, as well as other early work in Gricean semantics, such as Jonathan Bennett's (1976) *Linguistic Behaviour*) gives the impression that the target notion of speaker-meaning is the concept expressed by the ordinary language use of sentences of the form 'In uttering *x S* meant that such and such.' However, an attentive reading of Grice (or any other Gricean) reveals that that can't exactly be what is at issue. For example, Gricean accounts of speaker-meaning would approve the form of such speaker-meaning reports as:

(1) In uttering 'Is Frankfort the capital of Kentucky?' *S* meant *that you were to tell her if Frankfort was the capital of Kentucky*.
(2) In uttering 'Move to Frankfort, Kentucky!' *S* meant *that you were to move to Frankfort, Kentucky*.

Yet such reports are not likely to be found in ordinary speech. A speaker who uttered 'Is Frankfort the capital of Kentucky?' or 'Move to Frankfort, Kentucky!' would be said to have meant something, but if asked what the speaker meant, no ordinary speaker would answer with (1) or (2). Rather than respond with (1) and (2) she would be much more likely to respond, respectively, with 'He asked if Frankfort was the capital of Kentucky' and 'He told you to move Frankfort, Kentucky.' This doesn't mean that the Gricean is wrong to use sentences like (1) and (2) as canonical forms for reporting what speakers *mean* when they perform interrogative or imperatival speech acts, but it does seem to show that in using those representations he is stipulating a technical use of the verb 'to mean.' But then it's not clear what technical notion is being introduced. The Gricean's benchmark meaning reports are those he uses to report assertoric acts of speaker-meaning, such as 'In uttering "Il pleut" Pierre meant that it was raining in Paris.' In such reports the 'that'-clause is taken to refer to what the speaker meant, which is taken to be a proposition of some stripe or other. This suggests that the Gricean takes assertoric speaker-meaning to be a relation between speakers and propositions of some yet unspecified kind. At the same time, while the 'that'-clauses in (1) and (2) are being used to specify what the speaker meant, they can't be referring to propositions, nor, I should think, to anything else. One is left to wonder what univocal notion the Gricean takes speaker-meaning to be. Self-described Griceans today, such as Stephen Neale (see, e.g., Neale, 1992), are unconcerned that no Gricean analysis of speaker-meaning is recognized even by Griceans as being correct. Their commitment seems to be to the idea that what *S* means in uttering a sentence σ is determined entirely by intentions she has in uttering σ, where those intentions are intrinsically specifiable in non-semantic terms, and where the sentence uttered serves only as the means by which *S* intends those intentions to be recognized. I doubt that that idea can be correct. When Jack utters 'What's the capital of Kentucky?' his *intention* is to *ask* Jill what the capital of Kentucky is, and when Jill responds, 'Frankfort is the capital of Kentucky,' her intention is to *tell* Jack that Frankfort is the capital of Kentucky. I doubt that if we subtract those semantic intentions from all the intentions Jack and Jill had in

producing their utterances that there would be enough left over to entail that they performed the speech acts they in fact performed (Schiffer, 1987, ch. 9).

Other concerns emanate from *vagueness*. I think one salubrious effect of Gricean work on meaning is a better appreciation of the role of a speaker's intentions in the determination of what she meant and of what references she made, especially as such intentions may be crucial to the semantics of indexicals and other referring expressions. At the same time, I think certain features of vagueness show that the extent to which intentions determine what a speaker meant or referred to is quite limited. A single example should make clear what I have in mind. Jane is speaking on the phone with a friend in Los Angeles, and at one point she remarks, 'It's raining here.' Let's suppose that Jane made her remark while standing under an umbrella in pouring rain. Then her utterance is determinately true, and she clearly meant *something* in making her utterance. But *what* did she mean and what *determines* whatever it is that she meant? Well, what she meant is partly determined by the reference of her utterance of 'here,' so to what location did her utterance of 'here' refer? How one answers that question will depend on one's views about vagueness. The epistemic theorist of vagueness will say that there is some absolutely precise region of space α to which Jane's utterance of 'here' referred. If this is correct, then it should be obvious that that reference wasn't determined by Jane's referential intentions, since, having no way to distinguish α from the uncountably many precise regions with which it overlapped, she could not have intended to refer to α. The supervaluationist will say that there is no location that is determinately the referent of her utterance, but there are myriad – indeed, uncountably many – locations each of which is such that it's indeterminate whether it's the referent, or indeterminate whether it's indeterminate whether it's the referent, and so on (he will say that Jane's utterance is true because it's true in each of those "precisifications"). Suppose α is one of the uncountably many precise locations each of which is such that it's indeterminate whether it's the referent of Jane's utterance of 'here.' Then a theorist who wants to say that a speaker's referential intentions determine the reference of the indexicals she utters will want to say that it's indeterminate whether Jane intended to refer to α with her utterance of 'here.' But it should be obvious that she determinately did *not* intend to refer to α with her utterance of 'here' (cf. Buchanan and Ostertag, 2005, and Buchanan, 2010). How could she, when she couldn't pick out α from the billions of minutely differing overlapping areas if her life depended on it? A theorist who thinks that vagueness isn't confined to words and concepts might want to say that a certain vague area is the referent of Jane's utterance of 'here.' But if α is the vague area to which Jane's utterance of 'here' referred, Jane will be wholly ignorant of that fact: just ask her to tell you which locations are borderline – or borderline borderline, and so on – cases of being included in α. Of course, what goes for 'here' in Jane's utterance goes also, *mutatis mutandis*, for what she meant in uttering 'It's raining here.' The rub for intention-based semantics is that what goes for Jane's utterance goes also for virtually *every* utterance; for virtually every sentence uttered is to some extent vague.

What of the view that we must "explicate the intentional characteristics of language by reference to believing and to other psychological attitudes" (Chisholm, 1958, cited in Speaks, 2010), which seemed to motivate much of the work in IBS and arguably in Lewis's convention-based semantics.[18] It's difficult to say in the absence of a clear and plausible account of how exactly the intentionality of language derives from the intentionality of thought, and, most important, it's difficult to see why we should care whether the view is correct when what is apt to seem to be most important is that the intentionality of thought and language should supervene on non-intentional facts. At one time Brian Loar and

I thought it important to *identify* intentional facts with facts that are intrinsically specifiable in non-intentional terms, and we thought that to get such a reduction it was important first to reduce semantics to psychology, so that one could then reduce psychology to physical and functional facts in ways that seemed somewhat promising in the 1970s (see, e.g., Loar, 1981, and Schiffer, 1982). Alas, those ways no longer seem plausible (see Schiffer, 1987), not to mention that we no longer seem to have IBS at hand to reduce the semantic to the psychological.

Notes

1 Actually, I called the notion mutual knowledge* to make clear that I was stipulatively defining a technical notion, rather than trying to define a notion that was already called 'mutual knowledge' in the vernacular. It's not surprising Lewis and I independently hit on essentially the same notion, since we each invoked the notion to preclude essentially the same kind of counter-example, although for Lewis the counter-example he needed to prevent was to the account of convention he was developing, whereas the counter-example I needed to prevent was to the Gricean account of speaker-meaning I was trying to repair.

2 The generality *in sensu composito/diviso* distinction is from Abelard (Lewis, 1969, p. 64); those familiar with the knowledge *de dicto*/*de re* distinction will recognize that Abelard's distinction is definable in terms of that distinction.

3 There is a third problem I feel compelled to mention although I can't hope to develop it here (I do try to develop it in "How vagueness affects meaning," an unpublished manuscript that is still a work in progress). The problem is that a language can't be described as a pairing of sounds and meanings unless there are such things as meanings – that is to say, unless an expression's having meaning consists in there being some thing that it means. It is my view that that isn't what having meaning consists in.

4 Lewis (1975, p. 187) gives a pithy restatement of the problem: "A sentence never uttered at all is *a fortiori* never uttered untruthfully. So truthfulness-as-usual in [English] plus truthfulness-by-silence on the garbage sentences constitutes a kind of truthfulness in [Gobbledygook] … Therefore we have a prevailing regularity of truthfulness … in [Gobbledygook]. This regularity qualifies as a convention in *P* …."

5 Pinker (1994). Pinker attributes the example to Annie Senghas.

6 The notion of an illocutionary act – certain acts we perform in uttering sentences, such as telling, asking, requesting, ordering, warning, and so on – is due to Austin (1962).

7 For attempts to define illocutionary acts in terms of a Gricean notion of speaker-meaning, see Strawson (1964) and Schiffer (1972, ch. 4). For an attempt to define speaker-reference in terms of Gricean speaker-meaning, see Schiffer (1978; 1981).

8 Grice (1957, p. 385). Grice uses 'utterance' and its cognates in a technical sense that includes non-linguistic items and behavior.

9 There is an obvious problem for the Gricean account of assertoric speaker-meaning given that this is how recognition of intention is supposed to result in *A*'s believing the proposition *S* uttered *x* intending *A* to believe – namely, that while (3) might be a necessary condition for *S*'s *telling A p*, it's not a necessary condition for *S*'s *meaning p*; for if it were a necessary condition, then Grice would not have meant anything in writing any of the sentences in "Meaning," and this because, while Grice produced those sentences intending us to believe what they expressed, he certainly did not intend his readers to believe what he wrote on his authority; that is to say, he didn't intend his readers to believe anything he said because Grice's believing it was good evidence of its truth.

10 Relative to ongoing simplifying assumptions, γ is *simple signal* in population *P* just in case, for some q, γ means q in *P*, but there are no constituents of γ such that γ's meaning q is a function of the meanings of those constituents.

11 Cf. the account of the name-of relation in Evans (1973), which is based on the account of simple signals in Schiffer (1972).

12 An even better demonstration of the point being made is obtained by replacing the speaker-reference expressions in *N*, the plausible-looking definition of the name-of relation, with the definition of speaker-reference in terms of speaker-meaning I offered in (1981).

13 We saw in §1 that it's essential that the knowledge of *L* be *in sensu diviso*, lest speakers of a language be required to have something they clearly don't have – to wit, knowledge of the syntactic and semantic rules of *L*. Grice's own suggestion about how to define expression-meaning falls afoul of this requirement. His suggestion, put in the terms of this chapter, is that a sentence σ means *q* in *P* just in case for any (or nearly any) member *S* of *P*, *S* has a "resultant procedure" to utter σ if *S* wants to mean *q*, where a resultant procedure for σ would be a procedure that results from the procedures *S* has for uttering the constituents and structures of σ. Now that doesn't yet require *S* to have propositional knowledge of what his procedures are; the problem arises when we ask how another member of *P*, *A*, is to know what *S* means when she utters σ, for here Grice says that this is to be accomplished via *A*'s knowledge that *S* has in her repertoire the procedure of uttering σ if she wants to mean *q*. But *S* can hardly intend *A* to know that σ has the IB-feature of *being such that S has in her repertoire the resultant procedure of uttering σ if she wants to mean q* unless she intends *A* to compute that σ has that feature via her knowledge of the procedures *S* has in her repertoire for the morphemes and structures from which σ is composed, and it's impossible to see how that is to be accomplished without knowledge of the language's semantics and syntax. It will also be noticed that Grice's claim *that a sentence σ means q for S only if S has in her repertoire the resultant procedure of uttering σ if she wants to mean q* also runs afoul of the meaning-without-use problem.

14 An occurrence of a sentence is unembedded just in case the occurrence is not within some other sentence.

15 This seems very much along the lines of what Stephen Laurence (1996) has proposed as "a Chomskian alternative to convention-based semantics."

16 Much of this paragraph and the one before it is taken from Schiffer (2006).

17 This is elaborated in Schiffer (1982).

18 Avramides (1989) argues that Gricean semantics should be divorced from this motivation.

References

Austin, J. L. 1962. *How to Do Things with Words*, edited by J. O. Urmson. Oxford: Clarendon Press.

Avramides, A. 1989. *Meaning and Mind: An Examination of a Gricean Account of Language*. Cambridge, MA: MIT Press.

Buchanan, R. 2010. "A puzzle about meaning and communication." *Noûs*, 44(2): 340–371.

Buchanan, R., and G. Ostertag. 2005. "Has the problem of incompleteness rested on a mistake?" *Mind*, 114(456): 889–913.

Bennett, J. 1976. *Linguistic Behaviour*. Cambridge: Cambridge University Press.

Chisholm, R. 1958. "Chisholm-Sellars correspondence on intentionality." In Marres, 1958.

Chomsky, N. 2006. *Language and Mind*, 3rd edn. Cambridge: Cambridge University Press.

Evans, G. 1973. "The causal theory of names." *Proceedings of the Aristotelian Society*, suppl. vol. 47: 187–208.

Evans, G., and J. McDowell, eds. 1976. *Truth and Meaning: Essays in Semantics*. Oxford: Oxford University Press.

Gilbert, M. 1981. "Game theory and convention." *Synthese*, 46: 41–93.

Grice, H. P. 1957. "Meaning." *Philosophical Review*, 66(3): 377–388.

Gunderson, K. 1975. *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press.

Laurence, S. 1996. "A Chomskian alternative to convention-based semantics." *Mind*, 105(418): 269–301.
Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
Lewis, D. 1975. "Languages and language." In Gunderson, 1975, pp. 3–35.
Lewis, D. 1992. "Meaning without use: reply to Hawthorne." *Australasian Journal of Philosophy*, 70(1): 106–110.
Loar, B. 1976. "Two theories of meaning." In Evans and McDowell, 1976, pp. 138–161.
Loar, B. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.
Marres, A., ed. 1958. *Intentionality, Mind, and Language*. Champaign, IL: University of Illinois Press.
Neale, S. 1992. "Paul Grice and the philosophy of language." *Linguistics and Philosophy*, 15(5): 509–559.
Pinker, S. 1994. *The Language Instinct*. New York: W. Morrow.
Sawyer, S. ed. 2010. *New Waves in Philosophy of Language*. London: Palgrave Macmillan.
Schiffer, S. 1972. *Meaning*. Oxford: Oxford University Press.
Schiffer, S. 1978. "The basis of reference." *Philosophy of Language*, 13(1): 171–206.
Schiffer, S. 1981. "Indexicals and the theory of reference." *Synthese*, 49: 43–100.
Schiffer, S. 1982. "Intention-based semantics." *Notre Dame Journal of Formal Logic*, 23(2): 119–156.
Schiffer, S. 1987. *Remnants of Meaning*. Cambridge, MA: MIT Press.
Schiffer, S. 2003. *The Things We Mean*. Oxford: Oxford University Press.
Schiffer, S. 2006. "Two perspectives on knowledge of language." *Philosophical Issues*, 16(1): 275–287.
Speaks, J. 2010. "Introduction, transmission, and the foundations of meaning." In Sawyer, 2010, pp. 226–249.
Strawson, P. 1964. "Intention and convention in speech acts." *Philosophical Review*, 73(4): 439–460.

4

Meaning, Use, Verification

JOHN SKORUPSKI[1]

1 Meaning as Use

1.1 Introductory

Language has been the focus of the analytic tradition in twentieth-century philosophy. A good deal of that philosophizing about language has drawn its inspiration from a simple-sounding idea: to understand a word is to know how to use it. The formulation is particularly associated with Wittgenstein. But the idea itself has had immensely wide influence. It was important in logical empiricism – the empiricism of Vienna in the 1930s – and also in ordinary language philosophy in Oxford after World War II. It can be traced to the nineteenth century: for example, one might see it as a central feature of Peirce's pragmatist conception of meaning, or as a generalization on the reflections of philosophically minded mathematicians and scientists, in the latter part of the nineteenth century, about the meaning of scientific and mathematical calculi. (Notable among many were Mach, Poincaré, and Hilbert.) From the idea that use exhausts meaning important consequences have seemed to flow: the elimination of metaphysics, the dissolution of skeptical paradoxes – the pseudo-problematic nature of certain classical philosophical questions.

However, this chapter will not trace the nineteenth- and twentieth-century sources of the idea.[2] Nor will it examine the question of its grand philosophical implications, though these possible implications are of major importance. Our task here will be simply to assess the idea itself. We shall examine how it leads to a distinctive conception of meaning which I will call the 'epistemic conception' (§§1.3–1.5). Verificationism, an influential doctrine about meaning associated with the Vienna Circle, may be presented as a special case of this conception: §§2.1–2.3 will consider what verificationism is, its difficulties, and whether there can be a non-verificationist but still epistemic conception of meaning. In §§3.1–3.2 I will argue that important insights contained in the epistemic conception can be retained even if we treat them as insights about the normative nature of concepts rather than as

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

insights about the form of language-rules. And I will consider the effect of doing this on an influential doctrine whose modern form is closely associated with the epistemic conception of meaning – the doctrine that the *a priori* is the analytic.

1.2 Meaning and Use in Wittgenstein

"Meaning is use" says that use, function in a language, *completely* exhausts meaning. To understand an expression or sentence is to master its use within a grammatically structured means of communicating, that is, a language. No more is required for full understanding than whatever is required for that. But although this formulation is particularly associated with Wittgenstein, what he intended by it is a matter of controversy.[3] The invocation of use evoked a cluster of ideas, and commentators have highlighted different elements in this cluster.

Wittgenstein begins the *Philosophical Investigations* (2nd edn, 1958) with a critique of a conception of language according to which

> Every word has a meaning. This meaning is correlated with the word. It is the object for which the word stands. (§l, p. 2)

If we are mesmerized by the idea of meaning as a 'correlation' between a word and another thing, we misconceive what it is to understand a language.

Wittgenstein has many things to say against this mesmeric conception. He is particularly concerned to draw attention to the diversity of language uses, the variety of speech acts linguistic utterances can be used to perform – the many things you can do with words. Language is not just used to assert and describe. Nor are words just used to designate things. But we shall not be concerned with various important points he makes about the diversity of language use. (Baker and Hacker, 1980, provide a comprehensive commentary.) Our topic can be pinned down by distinguishing two criticisms of the correlational model. The first, widely made by many philosophers interested in language at least from Bentham onwards, is that certain expressions which seem to designate something may turn out, on analysis of their use in sentences in which they occur, not to do so. This is shown by producing a paraphrase of sentences in which the expressions occur, which preserves the meaning of the sentences but eliminates the expressions. This point does not put in question the correlational model of meaning as such.

The other point is more thoroughgoing and deeper. It is that the model of designator and thing designated is a philosophically misleading prototype of meaning. In making this point, one does not need to deny that a designation or 'semantic value' is associated with every ineliminable non-empty term, in virtue of its meaning. For example, the word 'yellow' will designate the property *yellow* – or a Fregean concept or the class of yellow things, or whatever the right account of its semantic value is – and we will understand that (inexplicitly) when we understand 'yellow.' But we can ask what it is to have that understanding. The *sentence,* 'The English word "yellow" designates the property *yellow*,' cannot be employed to explain the meaning of 'yellow' to someone who does not understand the word. Its meaning can be explained to someone who already understands another language, by using *that* language ('"Yellow" en anglais signifie *jaune*.'). But to someone who does not already understand another language it must be explained in other ways. Attending to the ways in which the meaning of words is actually explained

gives us an overview of their use and thus of the rules which govern that use. These rules constitute their meaning in the language.[4]

This line of argument for the conception of meaning as use will be a main topic in what follows. I will call it 'the Constitutive Argument,' since it is about what constitutes such metalinguistic knowledge as that 'yellow' in English designates yellow. We return to it in §1.5 and again in §§3.1–3.2. But finally in this section I want to note a point which is often connected with the Constitutive Argument in Wittgenstein's discussions of language – and also with his reflections on rule-following. In *Philosophical Remarks* (1975), for example, Wittgenstein says,

> in a certain sense, the use of language is something that cannot be taught, i.e., I cannot use language to teach it in the way in which language could be used to teach someone to play the piano. – And that of course is just another way of saying: I cannot use language to get outside language. (p. 54)

Now if the use of language is what cannot be taught, and meaning is use, one might conclude that meaning cannot be taught. But Wittgenstein only says that in a certain sense it cannot be taught. What does he mean? Any rule or instruction given for the use of a word must be given in language, understood broadly to cover all signs. Signs can only convey meaning if at some point there is a natural uptake of how they are being used. It is that natural uptake which cannot be taught – it is a condition of the possibility of teaching a language to someone that teacher and pupil share it. In this sense "Language must speak for itself" (Wittgenstein, 1974, p. 40). In grasping a language-rule, I grasp its applications – but this cannot require grasping further rules determining what its application to particular cases is. A being which grasps and applies rules must have spontaneous normative responses about the *right* way to apply a rule in a given case: responses not determined by a further rule. That normative dimension of understanding a sign cannot be conveyed by instruction in rules, but is presupposed by the very process of instruction.

We shall come back to this point as well in §§3.1–3.2. For the moment I simply note its compatibility with the previous one, which was that rules which constitute the meaning of a sign should be thought of as rules for its use. Nothing we have said so far precludes the possibility that such rules of use may be stated explicitly and systematically for a whole language, yielding thereby a theory of meaning for that language. Wittgenstein would probably have opposed such a view. But although it is true that rules of use presuppose normative responses which are not themselves codifiable as rules, that in no way shows that the rules do not exist or cannot be systematically exhibited.

So we turn now to the idea that meaning-rules are rules of *use*, rules for doing things with words.

1.3 The Priority Thesis and the Epistemic Conception of Meaning

Consider the speech act of assertion. It may or may not be the case that an account of it has to be given before we can give an account of other uses of language. It is at least clear that the assertoric use is a main use of language, for which there must be rules of use.

The most straightforward assumption to make about those rules would be that they combine to specify when a sentence in a language is correctly assertible. So where L is a language and S is a sentence in L, the specification has the form

(RA) S is correctly used to make an assertion in L if and only if …

Let us call the condition indicated by the dots on the right-hand side of 'if and only if' the *assertion condition*. Rules for the use of a word would contribute to determine assertion conditions for sentences containing that word.

In (RA) the notion of correctness is being used in a particular sense. It is correct, in the relevant sense, to use the sentence to make an assertion just if one is *justified* in making the assertion thereby conveyed – but questions of such matters as etiquette are not at issue. What is meant is that one is justified in thinking that assertion *true*. And the word 'true' is to be taken in its broadest sense, the sense in which any assertion whatsoever formally aims at truth. Truth in this sense may be partially characterized as a property F, such that for any assertion A whatsoever, if it is shown that there is no adequate ground to hold that A has F, reason (as against etiquette, discretion, etc.) requires withdrawal of A.

Why does the relevant notion of correctness relate in this way to the broad notion of truth? Because of what may be called the basic principle of the practice of assertion:

(A) One correctly uses a sentence to make an assertion if and only if one is justified in believing, of the proposition expressed by that use of the sentence, that it is true.[5]

Let us call this kind of correctness 'epistemic justification' – one uses a sentence correctly if one is epistemically justified in using it to make an assertion. So now we conclude that, in general, the assertion condition of a sentence will have to spell out its basic form of epistemic justification.[6]

In §2.3 we ask whether, even granted a conception of meaning as use, a sentence's meaning should be thought of as given *exhaustively* by its assertion conditions, or whether account should also be taken of the inferences it licenses. But perhaps our initial assumption – that rules fixing the assertoric use of a language should issue in direct specifications of the (RA) form – has been too speedy anyway? Look again at principle (A). It says that what one asserts and what one believes to be true is a proposition, not a sentence. If we reflect on that it may well strike us that the (RA) form over-ambitiously combines two tasks which should be kept separate. One task, that of the theory of meaning proper, is to specify for any sentence in L what proposition it expresses. (In the course of doing that in a finite systematic way, the theory will also have to specify for any expression in L what concept it expresses.) Another task is to give an account, for various kinds of propositions, of when one is justified in believing them to be true. This second task does not belong to the theory of meaning, but to epistemology.

Philosophers in the 1930s (Wittgenstein and the Vienna Circle) who took it that the way to specify the meaning of a sentence was (RA) also rejected this division between semantics and epistemology.[7] How are these theses connected?

A grounding idea is that *there is no language-independent account to be given of concepts and propositions*. To talk of concepts or propositions is simply to talk indirectly of the use of expressions and sentences in languages – classes of same-use expressions and sentences. Grasping a concept is understanding (the use of) an expression in a language. Grasping a proposition is understanding (the use of) a sentence in a language. Attitudes to propositions and concepts are attitudes to sentences and expressions in a language. We cannot *explain* understanding an expression or sentence as knowing what concept or proposition it expresses – as though that concept or proposition were an entity independent of language, and 'understanding what concept or proposition is expressed' were a matter of knowing the

correlation between the bits of language which do the expressing and the pre-existing non-linguistic item which is expressed.

Call this thesis the *priority thesis*.[8] It says that an account of concepts and concept-possession is dependent on an account of language-rules and language-understanding. It does not deny that it can be useful to talk of concepts and propositions. It is not denying the *truth* of principle (A). It is a positive thesis about what such talk amounts to. Talking about concepts and propositions is a way of talking about language-understanding, without specifying the particular language. It has a negative side – concepts and propositions have no explanatory role in the epistemology of understanding. We do not *explain* how a person understands the meaning of a word by saying that he or she possesses the concepts it expresses and knows that it expresses that concept. For possessing the concept just is knowing how to use the word (or some synonym) and that is what constitutes understanding it.

It is the priority thesis which seems to produce the conclusion that semantics and epistemology are one and the same. We may call this the *identity thesis*, for it denies that there are language-independent concepts which generate their own language-independent epistemic norms. There are only rules of language. Epistemic norms, the subject-matter of epistemology, are simply rules of classes of language – the subject-matter of semantics.

As I have noted, the slogan 'meaning is use' can be associated with a cluster of ideas in Wittgenstein's work, and its interpretation is controversial. It is clear that he himself directs it against the correlational model, and that he presents instead a conception of understanding as grasping language-rules which are like rules for making moves in a game. So he is not envisaging a reductive account of language-understanding in non-intensional terms, as some have thought (see, e.g., Horwich, 1995). To say that understanding consists in mastery of rules which are like rules of a game is still to give an intensional account – an account which attributes to language-users judgments about whether, for example, it is permissible or correct to utter a sentence. But we also thought it unlikely that Wittgenstein himself intended his emphasis on use to yield a systematic theory of meaning for a language. In order to leave that interpretative question clearly open, it will help to have a name other than 'use theory' for the view which does aspire to develop the idea of meaning as use systematically. Various names have been used – 'criterial semantics' (Baker, 1974; Peacocke, 1981), or 'anti-realist semantics' (Wright, 1987, ch. 7); 'verificationist semantics' (Putnam, 1983), or 'justificationist theory of meaning' (Dummett, 1993b). A first statement of this view is that ground-rules of meaning (for assertoric sentences) take the (RA) form, and that accounts of the meanings of words in a language must be given in such a way as to entail statements of that form for each assertoric sentence in L. Let us call it, non-committally, the *epistemic conception of meaning* (EM): 'conception,' in that it proposes what the *form* of a theory of meaning should be; 'epistemic,' because the meaning-rules it envisages state when assertion of a sentence is epistemically justified. This is only a first statement of EM: in §2.3, as I said, we shall consider the possibility of broadening it beyond this initial, verificationist form, letting it take into account what inferences assertion of a sentence justifies. But this form will do for the moment.[9]

The priority thesis seems to lead to the identity thesis and thus to EM. In identifying rules of language use with rules of epistemic justification it gives EM a particularly sharp and central role in philosophy. Just one story now gives a unified account of the meaning and the epistemology of L. The epistemic conception of meaning might just as well be called the semantic conception of epistemology. In effect it does away with the traditional philosophical discipline of epistemology. That does not make 'epistemic' a misleading word – it abolishes epistemology *because* it is an epistemic conception of meaning.[10]

This kind of view is central to logical empiricism. Logical empiricism held that there are only factual propositions – the province of science – and recommendations about how to speak or, more generally, what to do. There are no non-factual propositions, and there are no factual propositions which lie beyond the province of science. A language is a set of recommendations, or rules. The rules stipulate when a sentence in the language is assertible.

Think of the assertion condition as specifying an information state of the language-user. An important point is that every aspect of this state is accessible to the language-user. Whether or not one is in a state of experience, has a belief, has a justification for that belief – all this must be reflexively transparent if the rule is to be a rule of *use*. It must be possible in principle for me to tell, by reflection on my state of information alone, what it is and whether it warrants assertion of a sentence. If rules of use did not have this form I could not directly apply them: I would have to have a further criterion to tell whether the antecedent of the rule obtains.

Distinguish this from another point: must there be an effective procedure for deciding whether evidence warranting assertion of a sentence can be obtained or not? That is, must it always be possible to enlarge one's information state, by a specifiable method, to a point where one can authoritatively assert either that evidence warranting its assertion is available or that it is not? No. The requirement is that it should be transparent whether a sentence is assertible in an information state. *That* question must be effectively decidable. But the sentence itself need not in any sense be effectively decidable. One must be able to tell whether one's information state warrants the assertion 'There is evidence warranting assertion of S.'[11] If it does not, it's not required that one has a procedure for getting into an information state which decides the issue.

1.4 The Truth-Conditional Conception of Meaning

So the suggestion is that the priority thesis leads to EM, via the identity thesis. But we must now consider an analysis of meaning which seems to show that the suggestion is wrong. If this analysis is satisfactory, then the priority thesis does not entail the identity thesis, and therefore does not force EM.

The proposer of this analysis agrees that there must be rules which determine the correct assertoric use of sentences in L. But he insists on the point made in the previous section (§1.3) – those rules need only determine, for any particular sentence which can be used to make an assertion, *what* that assertion is. His claim is that we can formulate rules which do that, without making explanatory appeal to grasp of concepts and propositions, and without casting them in the (RA) form. They will be cast in such a way as to yield, instead, a statement for each assertoric sentence of the condition under which it expresses a truth (in the broad sense of truth invoked in §1.3). Such an account, he argues, tells us what each assertoric sentence says – and remains consistent with the priority thesis.

So instead of specifications of meaning of the (RA) form, this theory of meaning proposes to make do with specifications of the form

(RT) S is true in L if and only if *p*

('is true in L' means 'expresses a truth when used literally and assertorically in L').

Let us call this a *truth-conditional* theory of meaning. Like any other theory, it will need to make use of the compositionality of meaning: the fact that the meaning of a sentence is a function of the meaning of its constituent expressions. And it is this feature of the theory – its

appeal to compositionality – which is supposed to yield an account of understanding compatible with the priority thesis.

Consider for example the sentence, 'Ammonia smells.' Its meaning depends on the meaning of 'ammonia' and 'smells' and its syntactic structure. How might we spell out this dependence? Suppose the meaning of the words is given by 'dictionary' rules like this:

(1) 'Ammonia' is true (in English) of ammonia
(2) 'smells' is true (in English) of *x* if and only if *x* smells

And the syntactic structure is given by this 'compositional' rule:

(3) 'Fa' is true (in English) if and only if 'F' is true of that which 'a' is true of

Substituting 'ammonia' and 'smells' into (3) and using (1) and (2) we can deduce that

'Ammonia smells' is true (in English) if and only if ammonia smells.

So we know this equivalence solely on the basis of knowledge of those semantic rules of English in virtue of which (1)–(3) hold, plus the very basic logic involved in deriving it. That being so, the suggestion now goes, we know that the English sentence 'Ammonia smells' expresses the proposition that ammonia smells. For suppose that English sentence 'Ammonia smells' expresses a proposition P. Then I ought to be able tell that 'Ammonia smells' is true if and only if P is, just by knowing the semantic rules of English plus the basic metalogic of the theory. But let P = (say) the proposition that water is odorless. Then although it is true that ammonia smells if and only if water is odorless, I need to *know* that in order to recognize that

'Ammonia smells' is true in English if and only if water is odorless.

Although, in this example, the biconditional which I need to know is *a posteriori*, the point does not turn on that. Consider, for example, the proposition that $2 + 2 = 4$ if and only if $3 + 3 = 6$. This may be known *a priori*. But I still need to *know* it, as well as dictionary and compositional rules, to know that

'$2 + 2 = 4$' is true in English if and only if $3 + 3 = 6$.

In general, a sentence S in L expresses the proposition that *p* just if, by virtue of semantic conventions of L alone, S is true if and only if *p.* If I know an instance of this biconditional for every sentence in L which has an assertoric use, and I know it compositionally – through a grasp of the semantic value of the terms from which it is formed – then I have a complete grasp of L's assertoric power. So there seems to be no need to appeal to an account of the *assertion* conditions of sentences in L to account for the assertoric uses to which L may be put. The truth-conditional theory itself, so far as we have sketched it, seems consistent with the priority thesis. It does not mention concepts or propositions in explaining what it is to understand a language.[12] Moreover, it seems to provide a language-relative account of how one comes to know a proposition. Knowing the semantic conventions of L is knowing, for any sentence in L, what proposition it expresses. Which one? The one that is true solely on condition that that sentence expresses a truth. So grasping the proposition that ammonia

smells can consist in understanding English and then grasping it as the proposition which is true solely on condition that 'Ammonia smells' expresses a truth in English. (Or it can consist in understanding some other language L and grasping it as the proposition which is true solely on condition that the sentence in L which is in fact synonymous with 'Ammonia smells' expresses a truth.)

Thus the truth-conditional theory seems to be consistent with the priority thesis, in that it does not make explanatory appeal to the notion of language-independent concepts and propositions. On the other hand, it does not seem to require endorsement of the identity thesis either. It dovetails with the basic law of assertion, (A): I know that it is correct to assert a sentence S in L if and only if I have reason to think it expresses a truth – and the truth-conditional theory tells me what the truth-condition of S in L is. To know the *assertion* conditions of S in L I must both know its truth-condition, and also know the epistemology which links with that truth-condition. But that latter knowledge, knowledge of the appropriate epistemology, is not given by the truth-conditional theory of meaning itself. So the argument from the priority thesis to EM seems to break down.[13]

1.5 'Full-Bloodedness'

If it does break down, that must mean that the identity thesis is stronger than the priority thesis. It must mean that the priority thesis can be upheld consistently with a firm distinction between epistemic norms and rules of language. But can it be?

It is certainly true, as argued in the previous section, that if we know that

'Ammonia smells' is true (in English) if and only if ammonia smells

and we know that on the basis of knowledge of semantic conventions of English and very basic logic alone, then we know what proposition the sentence expresses – that is, that ammonia smells. But, as Michael Dummett (1974) has stressed, we can still ask what it *is* to know that 'Ammonia smells' is true (in English) if and only if ammonia smells. Call the proposition which is known – it is a metalinguistic proposition about English – 'M.' There is a difference between knowing M, and knowing that the metalinguistic *sentence* which expresses it is true. I could know that this sentence in the metalanguage (which in this case is itself English) expresses a truth without knowing what the object-language sentence meant, because I could know it to be true in English without grasping the proposition it expresses. (Compare: knowing that 'Lublin jest polskim miastem' expresses a truth in Polish, because you have been told authoritatively that it does, but not knowing what proposition it expresses.)

Can this point be deployed against the truth-conditional theory and in favor of EM? Is it an application of the Constitutive Argument (§1.2)? To deploy it in favor of EM one must take as one of its premises the priority thesis – which the truth-conditional theorist considered in §1.4 claimed to accept. The priority thesis says that to explain what it is to grasp a particular proposition is to give an account of what it is to understand some particular sentence or other. Now suppose we try to combine that with the claim that understanding 'Ammonia smells' is to be *explained* as consisting in a grasp of M. By the priority thesis, grasping M must then in turn be explained as consisting in understanding some sentence. What sentence? Well, we could say that grasp of M is explained by giving an account of what it is to understand 'Ammonia smells' itself – but that would now put us in an explanatory

circle. Apparently, then, we have to say that grasp of M is explained by giving an account of what it is to understand a sentence which expresses M. And then, by the same argument, we shall have to say that understanding that metalinguistic sentence will in turn be explained as grasping the higher-level metalinguistic proposition which specifies its truth-condition. Obviously, this won't do. It cannot be the case that every language is understood only by prior understanding of a metalanguage in which biconditionals about truth-conditions of sentences in the language are expressed.

But the choice between a vicious circle and a vicious regress arises from the attempt to combine the priority thesis with the claim that understanding a sentence is to be *explained* as consisting in a grasp of the metalinguistic proposition which specifies its truth-condition. Thus, if we accept the priority thesis we must reject that claim.

This spelling-out of the Constitutive Argument makes the priority thesis one of its premises. Similar reasoning forces rejection of the claim that understanding a word is to be *explained* as consisting in a grasp of a metalinguistic proposition – one which specifies its semantic value, or specifies the concept it expresses. Thus we are led to the conclusion that the theory of meaning must, in Dummett's words, be 'full-blooded' and not merely 'modest.' A modest theory of meaning, he says, is

> not intended to convey the concepts expressible in the object-language, but to convey an understanding of that language to one who already had those concepts.

while a full-blooded theory should,

> in the course of specifying what is required for a speaker to grasp the meaning of a given word, … explain what it is for him to possess the concept it expresses. (1993a, p. viii)

The point is that if we accept the priority thesis then we must reject the idea that understanding a word or a sentence can quite generally be explained as grasping a metalinguistic proposition which exhibits its meaning by specifying its semantic value or its truth-condition. On the contrary, we shall have to be able to say that grasping a metalinguistic proposition of that kind can consist in understanding the word or sentence which it is about. For example, grasping M can consist – if one's home language is English – in understanding 'Ammonia smells': the very same understanding as is involved, in that case, in grasping the proposition *ammonia smells*.[14] Explaining what it is to possess a concept or grasp a proposition becomes a task for the theory of meaning, and not for some other branch of philosophy. Hence there must be a part of the theory of meaning which does more than simply stating what expressions of the language are true of and deriving from that truth-conditions for sentences of the language. There *may* be a truth-conditional part of this kind, but there must also be a part which goes beyond it. And this part will conform to the epistemic conception of meaning.

But if we take this part to consist in the specification of assertion conditions for sentences in the language, won't the argument we have just considered apply to it as well? Won't it equally show that understanding 'Ammonia smells' cannot consist in knowing the proposition that 'Ammonia smells' is assertible iff … ?

To this the EM theorist's response is that knowing the assertion conditions of a sentence can consist in a *practical* ability to tell when it is right to utter it assertorically: to recognize information states as warranting or not warranting that kind of utterance of the sentence.[15] Precisely the same ability could, of course, be invoked to explain what it is to know the truth-conditions of a

sentence. But that is the EM theorist's point. To respond in this way would concede that grasping truth-conditions is not something over and above, independent of, mastery of assertion conditions. EM anchors understanding to a practical normative response.

Now all of this has proceeded on the assumption that the priority thesis is correct. But why cannot we reject that thesis, and accept an account of concepts and propositions which is not language-relative?

The most significant approach of this kind is *Platonism*. I use the term to refer to the view that concepts and propositions are non-spatio-temporal entities known by non-perceptual intuition. Platonism, combined with a truth-conditional view of meaning, may seem to offer an explanation of understanding. To know that 'straight' is true in English of straight things is to grasp, by non-perceptual intuition, the concept of straightness and to know that it is expressed by the English word 'straight.'

One can object, in this purported explanation of understanding, to the appeal to non-empirical intuition of concepts and propositions. But there is a different and clinching consideration – I will call it the 'no-intrinsic-meaning argument.' Wittgenstein uses it in various places, such as the following:

> In attacking the formalist conception of arithmetic, Frege says more or less this: these petty explanations of the signs are idle once we *understand* the signs. Understanding would be something like seeing the picture from which all the rules followed, or a picture that makes them all clear. But Frege does not seem to see that such a picture would itself be another sign, or a calculus to explain the written one to us. (1974, p. 40)

Wittgenstein's point is that there is no such thing as an object which has intrinsic meaning, that is, which (a) has meaning irrespective of having that meaning conferred on it and (b) is such that knowing it and knowing its meaning are one and the same. Even if we had access to objects in a Platonic third world, and had a mapping of terms and sentences onto these objects, that would do nothing for us unless those objects were already signs – signs which had intrinsic meaning. (If their meaning were not intrinsic, the questions of what it is for them to have meaning and what it is for us to understand that meaning would again arise.) The same would go for a picture in the world of physical or mental representations. The objection does not have to do with the particular world we are talking about. It is not a positivistic or even a naturalistic objection. (Blackburn, 1984, ch. 2 sets out a version of it and applies it to Fodor's 'language of thought' hypothesis.)

It is certainly a devastating argument against the view that a person's understanding of language is to be explained in terms of his or her possession of concepts and propositions – *if possession of concepts and propositions is taken to be quasi-perceptual access to a class of objects*. So taken, concept-possession could not *in principle* have a justificatory or explanatory role. But we have not shown that the only alternative to a language-relative account of concept-possession is one which treats concepts as intrinsically meaningful objects, mysteriously accessible to us. That would have to be shown, if we sought to derive the priority thesis from the no-intrinsic-meaning argument alone. Sometimes Wittgenstein seems to appeal to a dichotomy between an account of understanding which invokes access to intrinsically meaningful objects and one which invokes only grasp of language-rules:

> the mere fact that we have the expression 'the meaning' of a word is bound to lead us wrong: we are led to think that the rules are responsible to something not a rule, whereas they are responsible only to rules. (Reported in Moore, 1959, p. 258)

The apparent suggestion here is that if we avoid reifying 'the meaning' of a word into an intrinsically meaningful object then we have to accept that the rules governing its use constitute its meaning and are not 'responsible' to anything else. But may there not be a middle way – an account of concepts which neither reifies them nor makes them language-relative – and, given such an account, will it not be the case that a word which expresses a concept will have its meaning in the language set by rules which are 'responsible' to, or dovetail with, language-independent features of that concept? An account fitting this description would be this: to grasp a concept is to respond to a pattern of epistemic norms. It is to be disposed to accept a particular pattern of thought-transitions as primitively justified. Epistemic norms, however, are not themselves rules of language. A theory of meaning for a language is not in the business of describing them; that is a matter for the theory of concepts (or epistemology). Thus the theory of meaning can describe the rules of the language truth-conditionally, and will dovetail with an account of concepts which is neither language-relative nor Platonistic but characterizes possessing concepts as acknowledging patterns of epistemic norms.

Such an approach certainly has to reject the priority thesis, but it still accepts the no-intrinsic-meaning argument against Platonism. It provides, one might say, a full-blooded theory of concepts and a modest theory of meaning. So the question arises whether there is a case for the priority thesis which is independent of the no-intrinsic-meaning argument. We will return to these matters in §§3.1–3.2. But first we must examine further the conception which, as it now seems, is indeed forced if the priority thesis is accepted: that is, the epistemic conception of meaning.

2 Verificationism

2.1 Verificationism: Meaning and Truth

In the 1930s, verificationist conceptions of meaning were advanced by Wittgenstein and by philosophers of the Vienna Circle. But the connection between verificationism and EM is not straightforward. There can be non-epistemic versions of a verificationist view of meaning. And there can be non-verificationist forms of the epistemic conception of meaning. In considering these points we shall have to develop an account of EM which goes beyond the initial statement of it in §1.3.

I will use the term 'verificationism' to refer to a view of *meaning*, not, at least directly, to a view of truth. Verificationism is the view that understanding a sentence consists in grasping what information states would verify it. An information state verifies a sentence just if a person in that state is warranted in asserting it. All significant sentences have assertion conditions – their meaning can be displayed in the (RA) form.

In contrast, a verificationist view of *truth* holds that truth is verifiability. A sentence is true if and only if it is verifiable, that is, if and only if there is evidence warranting its assertion. To say that there is such evidence is to say, roughly, that a state of information warranting assertion of the sentence can be reached by us through an investigation which improves our current state of information – as it bears on the question of the sentence's truth or falsity – as much as it is actually possible to improve it.[16]

The difficulties with such a view of truth are notorious. Consider, for example, the two sentences 'Charlemagne's favorite color was magenta,' and 'Human beings cannot grow above 12 feet tall.' As far as the verificationist conception of *meaning* is concerned, both

sentences have a meaning. We know what kinds of evidence would warrant their assertion – for example, a text from the time of Charlemagne, which in general had the marks of reliability, and which recorded that Charlemagne often commented that his favorite color was magenta; inductive evidence that human beings never reach 12 feet together with theoretical considerations (e.g., relations between the height of an animal, its volume, mass, and muscular power, considerations of evolutionary fitness) which indicate that they could not. But we also know that evidence of that kind may not actually be available. It may not be possible to improve our information to the point where we are warranted either in asserting or in denying these sentences. Do we want to say that in that case those sentences are neither true nor false? Does a sentence's possession of truth-value depend on such contingencies?

To be sure, there are various ways of spelling out the word 'possible' in the phrase 'improving our information as much as it is possible to do.' Verificationists about truth characteristically idealize the notion of verifiability. For example, they may idealize the computing abilities of the agent which does the verifying, or its ability to move in space and time. But such idealized notions of verifiability cannot be identical with the concept of assertibility which is required for a verificationist view of meaning of the *epistemic* type. (This qualification will be explained in a moment.) For there the concept required, as was said in §1.3, is that of an assertion condition. And whether or not the assertion condition of a sentence obtains – whether or not the sentence is assertible in the language-user's information state – is something that must be transparent to the language-user. This transparent notion of assertibility cannot be identical with any non-transparent notion of verifiability – one which requires hypotheses about what would be assertible by an ideal agent.

There is, in fact, no straightforward route from a verificationist conception of meaning of this epistemic kind to a verificationist account of truth. It requires an unobvious philosophical argument to make the connection. (For arguments intended to make the connection see Dummett, 1959a; 1978b; Wright, 1987, "Introduction"; for criticism, see Skorupski, 1988.) On the other hand, there is a route from verificationism about truth to a non-epistemic type of verificationism about meaning.

A historical excursus will provide helpful background here. In conversations which he had in the late 1920s with Schlick and others from the Vienna Circle, Wittgenstein took a very strict verificationist line about meaning. The record made by Friedrich Waismann of these conversations contains many formulations of it – for example, 'The sense of a proposition is the method of its verification' (Wittgenstein, 1979, p. 79; cp. e.g., p. 227). Wittgenstein takes the notion of verifying a sentence quite strictly to mean '*indefeasibly* establishing its truth.' He describes (in Waismann's record) two conceptions of verification. According to one, the one he rejects, I cannot verify a proposition, for example 'Up there on the cupboard there is a book,' completely.

> A proposition always keeps a back-door open, as it were. Whatever we do, we are never sure that we were not mistaken.
>
> The other conception, the one I want to hold, says, 'No, if I can never verify the sense of a proposition completely, then I cannot have meant anything by the proposition either. Then the proposition signifies nothing whatsoever.'
>
> In order to determine the sense of a proposition, I should have to know a very specific procedure for when to count the proposition as verified. (Wittgenstein, 1979, p. 47)

Applying the procedure must yield a definite and indefeasible result. A consequence of this view is that general 'propositions,' which are not verifiable in the strict sense, have to be treated as 'hypotheses' rather than as genuine propositions.

But why must we adopt this very strict notion of verification? Why cannot verification just consist in achieving a state of information which warrants assertion? And why cannot that verifying state be defeasible? We know what kind of evidence would justify assertion of 'Charlemagne's favorite color was magenta,' or 'Human beings cannot grow above 12 feet tall.' We also know that that sort of evidence could be defeated by further evidence. We know that these sentences always 'keep a back door open,' that there can be no such thing as verifying them 'completely.' In short, why can't we work with 'defeasibly justify assertion of,' not 'conclusively establish the truth of'?

Whatever the reason for Wittgenstein's extremism in these Viennese discussions, his remarks usefully highlight the difference between two quite distinct philosophical perspectives from which verificationism can grow.

In one of these, it emerges from a combination of two things. The first is a conception of meaning which holds that a sentence has meaning by picturing a state of affairs (so, a species of the truth-conditional view). The second is an ontology which conceives of reality as a totality of states of affairs, thought of as immediately encounterable in experience. To understand a sentence is, then, to be able to picture – and for this kind of verificationism this means to be able to imagine experiencing or observing – the state of affairs which makes it true. And to verify a sentence or its negation is, so to speak, to run through the totality to the appropriate point and check by direct observation whether or not the state of affairs pictured by the sentence obtains. Verification, conceived in this way, is conclusive.

Here the central idea is that understanding a sentence is being able to represent to oneself what it would be like to encounter in experience the state of affairs which makes it true. Its affinity with Wittgenstein's *Tractatus* philosophy is suggestive. Though the Tractarian knowing subject is a highly elusive item, it is not implausible to think of it as being able to sweep at will through the states of affairs, or configurations of Tractarian objects, to which elementary sentences correspond, directly checking whether or not any elementary sentence is true.

But this last idea, with its phenomenalistic implication, could be loosened. The loosened version says that if one can describe, at least 'in principle,' what it would be like to have this encounter, the sentence is verifiable. It may not be possible to arrange to have the encounter, but the state of affairs is at least ideally verifiable – one can imagine a knower ideally transported to the site of the state of affairs and having the encounter. And now we have a verificationist notion of *truth* which can combine with a truth-conditional view of meaning to yield a kind of verificationism about meaning. Call this the positivistic route to verificationism. It rests on a positivist ontology of the real as the in-principle observable, and the verificationism which results is not a species of the epistemic conception of meaning.

But Wittgenstein does not say in these conversations that the meaning of a sentence is the picturable state of affairs which would verify it, render it true. His emphasis is on *methods* of verification. When he says on p. 227 (Wittgenstein, 1979) "The sense of a proposition is the method of its verification," he is quoted as continuing:

> A method of verification is not the means of establishing the truth of a proposition; it is the very sense of a proposition … To specify it is to specify the sense of a proposition. You cannot look for a method of verification. A proposition can only say what is established by the method of its verification.[17]

The next paragraph plays on the idea of thought as a movement with a determinate direction, set off in search of an answer to a question. The sense of both the question and the answer is given by the direction of the search (direction-sense).[18] Connectedly, Wittgenstein insists that different methods of verification ('thought-movements' with different directions) produce different senses. Such a view is not suggested by the positivistic route to verificationism. For as far as that conception goes I might be able to travel in various ways to the point of verification, the point at which I check by direct inspection whether or not the relevant state of affairs obtains. Equally, Wittgenstein's remark that you cannot look for a method of verifying a proposition – that is, first understand it and then look around for ways of verifying it – does not sit well with that positivistic conception.

Overall, then, it seems that what Wittgenstein presents in these conversations is a strictly operational kind of verificationism, a species of EM.[19] But without an underlying positivist (and indeed phenomenalist) impetus, there is no case for such strict operationism. The arguments for EM as such do not enforce it. Later Wittgenstein greatly broadened his operationist version of EM, taking into account the consequences for practice of an assertion as well as the operations which license it. To understand an assertion it is not enough to be told when you're licensed to make it: you need to know what it's a license to *do*, what consequences flow from it, ultimately for action.

The result is a fully liberalized, and pragmatized, conception of meaning as use. Understanding a word or sentence is knowing what can be done with it in communication and action, knowing the rules which govern its role in our practices of assertion and inference. The use of a sentence is as much a matter of the practical conclusions you can draw from an assertion of it as of the conditions under which you can assert it. The epistemic conception of meaning has now been framed in its full breadth. It is not derived *from* the idea that truth is verifiability. (The contemporary version of this line of thought, from verificationism about truth to verificationism about meaning, is to read 'is true' in a truth-conditional theory of meaning as equivalent to 'is assertible.' It is presented and discussed in Wright, 1987, chs 1, 2, and 9; see also Strawson, 1977.) Nor does it provide any obvious route *to* a link between verifiability and truth. In both cases the alternative, present in this tradition since the Viennese 1930s, is to endorse some deflationary view of truth. EM can realistically recognize that some – or, indeed, all – of the ways in which we acquire warrants for asserting a sentence are defeasible. An inquiry which was good enough to justify the assertion may be superseded by further inquiry which defeats that assertion, that is, leads to an improved information state in which the assertion is no longer justified. The epistemic conception, comprehensively stated, is compatible with such defeasibility in a way that the strict verificationism enunciated in the passage from Wittgenstein quoted above is not.

In fact if we adopt this comprehensive epistemic conception of meaning (for short I will call it 'the comprehensive EM') we have to reject not only strict verificationism but verificationism as such. For to say that evidence defeasibly warrants an assertion that *p* is to accept as intelligible 'There is evidence warranting the assertion that *p* but it is not the case that *p*.' This sentence must have meaning since it appears as a constituent in 'It is logically possible that there is evidence warranting the assertion that *p* but it is not the case that *p*.' The latter sentence is one which we are justified in asserting if we are justified in holding that there can be evidence that *p* which is sufficient but defeasible. It is the way we express, in the language, the proposition that evidence is defeasible. Yet the constituent sentence itself is never verifiable. Thus, if a sentential constituent of a meaningful sentence must itself be

meaningful, we have a sentence in the language which is meaningful but not verifiable. So we cannot liberalize verificationism to allow for defeasible verifications: if we liberalize it we have to go beyond it.

We return to this point in §2.3. But first we will consider what account the comprehensive EM can give of the meaning of the logical operators. They are of great importance – witness the fact that sentences like the one above, whose intelligibility refutes verificationism, contain them.

2.2 The Meaning of the Logical Operators

We can give an account of the meaning of logical operators (the operators of sentential logic, and the quantifiers of predicate logic), as of any other expressions, contextually: by giving an account of the way they contribute to the meaning of sentences in which they occur. But on the verificationist view of meaning, an account of the meaning of complex sentences containing the logical operators will have to take the RA form. So our account of the meaning of logical operators, on the verificationist view, must spell out how they contribute to the assertion conditions of sentences in which they occur. And it is natural to think that the way in which they contribute is by mapping the assertion conditions of the constituent clauses of the complex sentence onto an assertion condition for the complex sentence itself; just as in a truth-conditional theory they map the truth-conditions of the constituent clauses onto a truth-condition for the complex sentence itself. Call this the 'assertion-condition-functional' (ACF) view of their meaning, as opposed to the truth-condition-functional (TCF) view of their meaning advanced by the truth-conditional approach. As Dummett famously put it:

> *We no longer explain the sense of a statement by stipulating its truth-value in terms of the truth-values of its constituents, but by stipulating when it may be asserted in terms of the conditions under which its constituents may be asserted.* (Dummett, 1959a, pp. 17–18 in 1978a. Emphasis in the original)

The ACF view leads to the conclusion that verificationism will require rejection of classical logic. For consider 'P or it is not the case that P,' and compare its truth-condition – it is true if either 'P' is true or 'It is not the case that P' is true – with the assertion condition it will have on Dummett's proposal: it is assertible if either 'P' is assertible or 'It is not the case that P' is assertible. This latter account of the meaning of 'or' will allow us to assert 'Either magenta was Charlemagne's favorite color or it was not' *only* if we have evidence warranting the assertion that it was or evidence warranting the assertion that it wasn't. Classical logic, on the other hand, allows us to assert the sentence outright. To save classical logic, one might try supplementing one's account of the meaning of 'or.' For example, we could say that a sentence of the form '*p* or *q*' would be assertible just where '*p*' is assertible, or '*q*' is assertible, *or where* '*q*' = '*not-p.*'

There are serious obstacles to this suggestion; but they need not concern us.[20] For, whatever one's view may be about the desirability or otherwise of maintaining classical logic, there is something wrong with the idea that an account of the logical operators must be ACF. The point turns on this: evidence that there is no evidence that *p* does not warrant asserting that it is not the case that *p*. For example, we may have sufficient warrant to assert that there is no evidence that Charlemagne's favorite color was magenta; but that does not

justify us in denying that Charlemagne's favorite color was magenta. Now consider a pair of sentences of the form '*p*' and 'It is assertible that *p*.' They have the same assertion conditions; any information state which warrants assertion of the one warrants assertion of the other. Consider next the pair 'It is not the case that *p*' and 'It is not assertible that *p*.' These clearly do not have the same assertion conditions, for the reasons just given. It follows that an ACF account of the meaning of 'not' cannot be acceptable.

On any view, it is not part of our practice to regard a demonstration that there is no evidence that *p* as tantamount to a warrant for asserting 'It is not the case that *p*.' The reason is obvious. The world is not totally surveyable by us. There are true propositions about it which we do not have the evidence to assert. Evidence can sometimes be sufficient, though defeasible; but it can also be simply insufficient.

So, also, 'If *p* then *q*' cannot mean, for example, 'If it is verifiable that *p* then it is verifiable that *q*.' For let '*q*' = 'there is no evidence that *p*.' The sentence, 'If *p* then there is no evidence that *p*' is perfectly intelligible and may indeed be assertible. ('If the Prime Minister is a master-criminal there is no evidence that he is. For a master criminal is totally effective in covering his traces.')[21] Both '*p*' and 'there is evidence that *p*' have assertion conditions, but the assertion conditions of the conditional cannot be a function of them.

But does the comprehensive EM have to give an ACF account of 'not' and 'if'? Well, there is no ban on its using the word 'true' in formulating assertion conditions for complex sentences. For it does not deny that a person who understands a sentence S in L can thereby be said to know that S is true in L if and only if *p* (where the sentence which replaces '*p*' has the same semantic content as S). On the contrary, it says that knowing that metalinguistic proposition just is understanding S (see §1.5) – and understanding S in turn, according to a comprehensive EM, consists in mastery of its use in the language – of when it is assertible and what can be inferred from it.

So a comprehensive EM does not require that accounts of the meaning of logical operators must be ACF. It can allow that users of L who understand, and thus grasp the truth-conditions of, elementary sentences in L (those not containing the operators) may also have a truth-functional understanding of the operators. This view, incidentally, does not require that the meaning of those elementary sentences is unaffected by the introduction of operators into L. If introducing an operator into L changes the inferential power of a sentence which does not contain that operator, it also changes its use in L and thus its meaning. But this does not offend the compositional principle, that the meaning of a sentence is determined by the meaning of its constituents. That principle does not preclude the possibility that introducing a new operator into L changes the meaning of sentences in L. It only says that the meaning of a sentence formed with the new operator is a function of the meaning of its constituent sentences.

Let me illustrate by reference to the word 'not.' A truth-functional specification of its meaning will say: 'It is not the case that *p*' is true if and only if '*p*' is not true. If I know that, I can infer that

'It is not the case that *p*' is assertible if and only if it is assertible that 'p' is not true.

So I know the assertion condition of 'It is not the case that *p*' – so long as I can recognize the conditions which warrant denial of '*p*' (i.e., assertion that '*p*' is not true). But I cannot, from the assertion condition of a given sentence, mechanically derive the assertion condition for its denial, and hence not the condition for assertion of its negation either. So this is an

account of the assertion condition of 'It is not the case that *p*' which is not ACF. Nothing in the comprehensive EM requires that a semantics for the word 'not' should equip me with the ability to recognize when denial of any arbitrary English sentence is justified, solely as a function of its assertion condition. A semantic theory for English tells me that the correct way to negate an English sentence '*p*' is by saying 'It is not the case that *p*.' It registers the semantic complexity of negations by delivering truth-conditions for negations as a function of truth-conditions of the sentences negated. In doing so it enshrines the substantive principle that negation of S is justified just if denial of S is: a fundamental normative feature of our inferential practice. But semantic theory has no mission to tell me any more than that. Of course, on the epistemic conception of meaning there must still in principle be an account of one's understanding of the assertion conditions for denial of various kinds of sentence. But these may be very multifarious and will not be functions of the assertion conditions for '"P" is true.' The same goes for the other truth-functional operators.[22]

2.3 Beyond Verificationism

We noted in §2.1 that verificationism cannot be liberalized without being rejected. Liberalization means recognizing that the best available evidence may (a) be insufficient and (b) when sufficient at present, may yet in future be defeated.

For example we are entirely justified in saying that Charlemagne's favorite color may have been magenta even if there is no evidence warranting the assertion that it was or was not. And we are also justified in saying that, while there are currently sufficiently good scientific grounds for thinking that nothing can travel faster than light, it remains possible that theoretical advances in future may defeat them.

More generally, we are justified in holding that (1) there are sentences which are true even though no one has sufficient evidence for asserting that they are, and (2) that there are sentences which we have sufficient evidence to assert, but which are not true. Both these general propositions, about our ignorance and fallibility, are justified as internal consequences of our overall commonsense and scientific conception of the physical world (of which classical logic is currently a part), our place in it, the way we get causal signals from it, and so on. The intelligibility and truth of (1) and (2) cannot be denied; but at the same time there can be no warrant for asserting any of their instances.

So those instances, for example '"Charlemagne's favorite color was magenta" is true but there is insufficient evidence to assert "Charlemagne's favorite color was magenta,"' or 'We are justified in asserting "Nothing can travel faster than light" but that sentence is not true,' have no free-standing assertoric role in the language. There are no circumstances which justify their assertion. Their only role is as constituents in complex assertions, embedded, for example, in the context "It is possible that …," "It could be true that …," or in conditionals or negations. But it is still a role; and it is a grasp of that role which, according to the comprehensive EM, constitutes our mastery of their meaning.

The truth in verificationism is that where a sentence can have free-standing assertoric use, grasp of its meaning requires mastery of that use. But some complex sentences formed by sentential operators have a meaning only in virtue of their role in inference and their embedding in more complex sentences still. Where a sentence has a free-standing assertoric use, a person who understands it will know that it has that use and thus will have a grasp of its assertion conditions. But where a sentence does not have such a use, but can still figure in embeddings and inferences, that is what is grasped by someone who understands it.

3 Rules and Norms

3.1 Concepts as Cognitive Roles

In §§2.1–2.3 we attempted to set out EM in its broadest, most plausible, form. But we must now go back to the questions raised at the end of §1.5. That section argued that the priority thesis amounts to the identity thesis and imposes an epistemic conception of meaning or, otherwise put, a semantic conception of epistemology. We saw how Wittgenstein's no-intrinsic-meaning argument destroys any conception of meanings or concepts as intrinsic signs. But we also saw that the no-intrinsic-meaning argument seems to fall short of establishing the priority thesis and thus EM.

Grasping the meaning of a word cannot consist in cognizing an intrinsically meaningful object. But this does not refute the simple point that to understand a word *is* to possess a concept and know that the word expresses the concept. It merely shows that possessing the concept is not a matter of cognizing any such object. The right response to Wittgenstein's no-intrinsic-meaning argument may be a better account of concept-possession than that of the Platonist (or the 'language-of-thought' theorist).

A better account is that to possess a concept is to acknowledge certain cognitive moves as justified. Grasping concepts is acknowledging norms. By analogy to the slogan that meaning is use, one may say that *concepts are cognitive roles*. The no-intrinsic-meaning argument does not decide the choice between the two slogans.

Are the slogans complementary, or does one make the other redundant? It is a question of the difference between norms and language-rules. By a 'norm' I mean a true normative proposition about reasons. An *epistemic* norm is about reasons to believe – about the relation '… gives *x* reason to believe that *p*.' So the slogan 'Concepts are cognitive roles' says that to possess a concept is to acknowledge a pattern of epistemic norms. In contrast, a rule is not a *proposition* at all. It cannot be said to be true or false. It is the content of an explicit stipulation or implicit convention. The priority thesis comes down to saying that we cannot treat purported epistemic norms as ultimately distinct from *rules* of a language. Talk of norms constituting a concept must reduce to talk of language-rules constituting the meanings of words.

The no-intrinsic-meaning argument does not establish this thesis. We must look elsewhere – to that extraordinarily influential assumption which (as we noted in §1.3, in discussing the priority and identity theses) was made by Viennese logical empiricism. It was also made by Oxford ordinary-language philosophy, and indeed Quinean naturalism. The assumption is that all propositions are factual. Assertoric and judgable content is factual content.[23] In that case, if there are normative propositions there must be a domain of 'normative facts.' Well, we do talk about 'the fact that' one ought to come to the assistance of distressed people, or 'the fact that' one ought to accept the simplest explanation of the data. But we are not, I think, indulging in ontology. There is a substantial, ontologically committing use of the word 'fact': in this use of the word the idea of 'normative fact' seems to be a kind of category mistake. The stubborn thought that makes it seem a category mistake is a cousin, one might say, of the no-intrinsic-meaning argument. It is the thought that no fact, in any world (natural or non-natural), is intrinsically normative. Acknowledging a norm cannot *consist* in recognizing a fact. Norms are no more facts than meanings are things. But it is in this ontologically committing sense of 'fact' that the claim that all propositions are factual is to be understood.

If all propositions are factual and there are no normative facts, normative utterances, such as 'You ought to come to the assistance of distressed people,' or 'You ought to accept the simplest explanation of the data,' cannot be assertions but must rather be understood as recommendations, proposals, prescriptions, and so forth. In particular, then, it can become plausible to hold that the alleged epistemic norms which constitute concepts should really be seen as prescriptions as to the use of words. But this conclusion is not enforced by the powerful double-barreled weapon that says no object is intrinsically meaningful and no fact intrinsically normative. It requires the further claim that all assertoric content is factual. Only then do we get the dichotomy of facts and rules which generates the priority thesis and EM.

Although Wittgenstein often seems to assume the dichotomy (as in the second passage quoted on p. 82) it is also Wittgenstein, especially in his later thinking, who effectively drives a wedge through it in his reflections on what it is to follow a rule. He highlights the point, noted in §1.2, that to apply a rule is to exercise normative judgment. What view he takes of it, having highlighted it, is a matter of dispute. Here I am assuming, contrary to some readings of his philosophy of language, that he does not intend to *deny* that the question, 'Has the rule been applied correctly?' can have a true answer. His view is not the nihilist one that there is no true answer, or the extreme-conventionalist one that the answer in every case expresses a decision. He accepts that it can be determinately true that if you're following the rules of English you ought to call this patch here 'yellow' (though there can also be vague or indeterminate cases). I also assume – contrary, admittedly, to much current discussion – that Wittgenstein was not a reductionist. It was not his view that 'If you're following the rules of English you ought to call this patch here "yellow"' has a non-normative truth-condition, consisting, say, in a fact about the speech-dispositions or mental states of certain language-users. But if nihilism, radical conventionalism, and reductionism are all false then we have here an example of a normative judgment which corresponds to no fact (in the ontologically committing sense).[24] The upshot is that a thinker who follows rules must grasp norms *as well as* facts and rules. Commitment to the existence of norms is thus entailed by our very description of an entity as a rule-follower – if there are rules, the dichotomy of facts and rules is not exhaustive.

Applying a rule involves a spontaneous normative capacity which is reducible neither to judgments about what is the case nor to familiarity with conventions or stipulations. But why should interpretative normative judgments, judgments about the right way to apply a rule to a case, be the *only* instances of true normative propositions? We naturally and stably converge on many primitive judgments about what there is reason to think, feel, or do. Spontaneity and stability of normative judgment is present in all these cases. They are genuine judgments; no more is needed to show they have genuine propositional content.

Now we can formulate a real contrast between an epistemic conception of meaning and an epistemic conception of content (or concepts). Both hold that a truth-conditional theory of meaning must be supplemented if one wants a full account of language-understanding. And both can be said to hold that the supplement must be an account of concepts as cognitive roles. But EM takes it that an account of the cognitive roles of concepts reduces to an account of rules for use of expressions in a language. It holds that the required supplement is still *semantic*. On this view, there is a level of semantic theory which describes conventions for introducing and eliminating terms in a language. Conventions stipulating when a sentence is assertible and what is inferable from it are determined by them. They constitute the language, and the level of semantic theory at which they are stated – call it 'the cognitive-role level' – is more fundamental than the truth-conditional level. In contrast, an epistemic

conception of content ('EC' for short) takes the objectivity of norms seriously, and holds that an account of concepts can consist in an account of the epistemic norms regulating their introduction and elimination in one's thinking. Such an account – a theory of epistemic norms – is not a level of semantic theory, for it does not purport to describe rules of a language. It denies the identity of semantics and epistemology. As far as the semantics of a language is concerned, it can hold that a truth-conditional account is fully adequate.[25]

Many questions are raised by this approach; a number of them are analogous to questions which arise for a comprehensive EM. Thus one can ask how a theory of epistemic norms, as well as a theory of linguistic rules of use, copes with the phenomenon of defeasibility; and one can ask how concept-constituting norms, or rules, determine an extension for a concept. These are crucial questions, but they will not be pursued here. The next and final section takes up a very important and attractive corollary of EM: the account it yields of how *a priori* knowledge derives from grasp of meaning.

3.2 Aprioricity and Normativity

An epistemic conception of meaning greatly enlarges the empiricist idea that aprioricity is analyticity – that an *a priori* warrant for an assertion is one obtainable from a grasp of its meaning alone. Because it introduces rules of language at the cognitive-role level it is able to give a new account of analyticity which differs from what one might call the Kant/Mill account.

In the latter, a class of sentences is identified as uncontroversially empty of content, or a class of inferences as uncontroversially 'merely apparent'; and then these, together with sentences or inferences reducible to them by explicit definitions, are defined as analytically true. Take, for example, 'Anyone who is a father is a parent' or, 'He's a father. Therefore he's a parent.' The explicit definition is '"*x* is a father" = Df "*x* is male and *x* is a parent."' The contentless sentence might be 'A father is a father,' and the inferences acknowledged as merely apparent would in this case include and-elimination. But as Mill particularly emphasized, this account of analyticity does not guarantee that *all* logic is analytic. It is not uncontroversial that all logically valid inferences are merely apparent, even if it is uncontroversial that and-elimination is. (If even this is rejected, the class of analytic truths is even smaller: e.g., 'Tomorrow is the day after today.') In this respect, the Kant/Mill account contrasts with the 'Kant/Frege' account, which characterizes analyticity outright as derivability, with explicit definitions, from logic. However, it does not (in Frege's case at least) claim that analyticity is truth by virtue of meaning alone, or that analytic propositions are empty of content; and it is therefore unacceptable to a clear-headed empiricist. In contrast to both of these approaches, then, the new account of the *a priori* generated by EM *does* simultaneously claim that all logic is analytic and that analyticity is truth by virtue of meaning alone. It promises an empiricist account of the aprioricity of logic and mathematics. This has been perhaps its most influential feature.[26]

In the new account, as in the Kant/Mill account, a sentence is *a priori* or analytic when a justification for asserting it can be derived exclusively from a grasp of its meaning. But the rules which constitute that meaning will now include introduction and elimination rules statable only at the cognitive-role level. An example will explain what I mean. Consider the following introduction rule for the English word 'yellow':

(1) The occurrence of a visual experience as of a yellow object in one's visual field warrants, in the absence of defeating information, assertion of the sentence in English 'There's something yellow there.'

This is a rule formulated at the cognitive-role level. In contrast, if (as is plausible) 'yellow' is semantically simple, then a truth-conditional semantics for English will contain only the following dictionary rule:

(2) '"yellow" is true (in English) of *x* if and only if *x* is yellow.'

Notice that as I have formulated (1) the relation *warrants assertion of* holds between a state of visual experience – something which is not a sentence – and a sentence. Many philosophers, both friends and foes of EM, would find this unacceptable. They assume or argue that the relation can only hold between sentences.[27] But the only relevant constraint on an object which satisfies '… warrants assertion of S,' where 'S' can be any sentence, seems to be that it must have content and be transparent (in the sense of §1.3). A visual experience or a memory has content and is transparent, and so it satisfies that constraint. Some rules at the cognitive-role level will link warrants for asserting sentences to warrants for asserting other sentences (specifically, in the case of logical connectives, metalinguistic sentences – see §2.2). But if a language has empirical content at all it must contain rules linking the assertibility of certain sentences in the language to the language-user's experience and memory. In the spirit of stating EM in the most liberal way possible, we should allow that a fully comprehensive EM account can include them.

Consider now the following normative proposition:

(3) The occurrence of a visual experience as of a yellow object in one's field of vision justifies, in the absence of defeating information, a judgment that there's something yellow there.

This is not a metalinguistic statement of a rule of English as (1) is, but a normative proposition stated in English. What is the relation between them? The EM theorist must maintain that (3) is in some way an expression of (1) alone. It cannot be a genuine normative proposition. Rather, sentence (3) is 'assertible *a priori*' in English because its warrant derives solely from a rule of the language – for English the rule will be (1), while for other languages which have a sentence synonymous to (3) it will be a rule analogous to (1). Let us allow, for the sake of argument, that the details of this can be filled in coherently. However it is done, the crucial point is that it provides an explanation of how *a priori* knowledge of (3) is possible, in a way that no appeal to (2) could do. I have *a priori* knowledge of (3) in virtue of grasping rule (1), or some analogous rule in another language.

In short, EM generates a new account of aprioricity as analyticity, because it postulates introduction and elimination rules at the cognitive-role level. Indisputably, this is important and new – a major twentieth-century contribution to philosophy. But is it right? *Does* the aprioricity of (3) depend in any way on there being a rule of English expressible by (1), or some analogous rule for another language? Well, it's far from obvious that it does depend on that. Do we want to say that (3) is '*a priori*'? What makes us want to say it is, if we do, does not seem to stem from the fact that some language, English or another, contains some rule. Rather, the essential point is simply that we converge, on critical reflection, in finding (3) primitively or spontaneously compelling. It is a fundamental epistemic norm: it expresses a primitive normative response. Acknowledging it does not consist in learning any linguistic convention or stipulation. It's the other way round: training in linguistic conventions *assumes* such primitive normative responses. In teaching (2) we assume the existence of belief-forming dispositions responsive to (3). And once a person had learned (2) he or she would see the truth of (1), understood now not as a language rule but as a consequence of (2) and (3).

On the other hand, we have accepted the point that no fact is intrinsically normative. So acknowledging (3) does not consist in knowledge of any fact, natural or non-natural. What, then, is its epistemology? As for any other fundamental norm, be it of belief, action, or feeling, it is the epistemology of reflective examination and critical convergence. That is the epistemology characteristic of the normative: it is not the epistemology appropriate to propositions which depict the existence of a state of affairs.

This takes us to the brink of controversial epistemological questions which are not on our agenda here. For present purposes it is enough to pin down how EC, the view that concepts are patterns of epistemic norms, differs from EM, the view that they are patterns of language-rules. EC requires the thesis that the normative and the factual are both domains of judgment, consisting of propositions with truth-value. If this is defensible, then we can say that (3) expresses a norm partially constitutive of the concept *yellow*. We can also say it is a 'conceptual truth' – at any rate it is concept-constituting and it is true. *But its status as a conceptual truth in this sense in no way explains how it might be* a priori. The way it is known to be true is the way that any fundamental norm is known to be true; its epistemology is that appropriate to fundamental norms in general. It is not *because* it is a conceptual truth, constitutive of the concept *yellow*, that it is true. Thus EC does not belong to that class of views which takes certain truths to be '*a priori*' and seeks to *explain* that status by saying that they are conceptual (truths, that is, which go beyond the Kant/Mill prototype of analyticity). But this does not matter. If concepts are constituted by norms of reasoning, and if we can get a satisfactory account of normative knowledge, we do not *also* need a substantive theory of the *a priori* which goes beyond Kant/Mill analyticity. An account of normative knowledge will do what an account of the *a priori* was meant to do.[28]

I assumed earlier (§3.1) that Wittgenstein is neither a nihilist nor a radical conventionalist nor a reductionist about rule-following. If all this is right then the later Wittgenstein needs a distinction between rules and norms of the kind made here. His own reflections on rule-following show that to avoid this trilemma one must go beyond the Viennese dichotomy of facts and rules. It is that dichotomy, together with the points that no object has intrinsic meaning and no fact is intrinsically normative, that produces the package of EM, the linguistic theory of the *a priori*, and radical conventionalism about logic and rule-following. But did Wittgenstein go beyond it? Reading his later writings on 'grammar' and 'rules' it is hard to come up with an answer. Michael Dummett attributes the whole package to Wittgenstein (see, e.g., Dummett, 1959a, 1994; cp. Stroud, 1965). Others disagree. They faithfully reflect Wittgenstein's own murkiness. In a valuable discussion of Wittgenstein's notion of a criterion, for example, Hacker (1990) comments thus:

> To say that *q* is a criterion for W is to give a partial explanation of the meaning of 'W,' and in that sense to give a rule for its correct use. The fact that the criterial relation between *q* and W may be neither arbitrary (in one sense at least) nor stipulated, that in innumerable cases we could not resolve to abandon the normative relationship without a change in our form of life, and in many cases could not abandon it at all, does not imply that it is empirical, let alone that it is a matter of *Wesensschau*. We may concede that certain concepts are deeply embedded in our lives, occupy a pivotal role in our thought and experience, yet still insist that their use is rule-governed, a matter of *nomos* rather than *phusis*. (p. 552)

Note how the line of thought here goes from acknowledging that the relationship is 'normative' to the conclusion that – however inescapable for us, however felt as a constraint rather than a stipulation – it must yet be a 'rule,' a matter of convention rather than nature. But why cannot it be acknowledged that it is normative without being in *any*

sense a convention? It is hard to see what could be at work here other than the philosophical thesis that all propositions, judgable contents, are factual.

It may be impossible to tell how far Wittgenstein thought his way past this thesis. On the one hand, he was not (in his later thought) burdened by the realist semantic assumptions about truth and reference which lead to it. But on the other hand, his constant insistence that 'training' determines the 'logical grammar,' or framework, of our language-games at least suggests that he did not repudiate the Viennese dichotomy of facts and rules. For one is 'trained' to observe rules: the process of acknowledging a norm – spontaneously, autonomously – is a process of education, not 'training.'

At any rate, if we reject the thesis that all judgable content is factual, we can acknowledge that the normative is a domain of the understanding, something we judge of – but yet that norms are still like rules in this respect: we do not find them in the world. They are presupposed in cognition of a world – and that view still has certain strong affinities with Wittgenstein's later philosophy, even if it is not his. For example, what he says about logic would also apply to this view of norms. For them, as for 'logic,'

> There is not any question at all … of some correspondence between what is said and reality; rather is logic antecedent to any such correspondence. (Wittgenstein, 1978, l.156, 96)

Certainly this a very important and controversial philosophical claim. The question in the end, of course, is not who thought it but whether it is true, and if so, how or why.

Notes

1 I am grateful to Bob Hale and Crispin Wright for many very useful discussions (including some illuminating disagreements) about the issues dealt with in this chapter. The points made in §3 are developed more fully in Skorupski (2010).

2 I shall refer to its development in Wittgenstein's thought, since his discussions of it remain influential and exemplary. A balanced historical account would also examine the important ideas of a number of his contemporaries; for example, Rudolf Carnap and Moritz Schlick in the development of Viennese verificationism and Friedrich Waismann for his influence on the development of ordinary language philosophy (e.g., Carnap, 1936; 1937; 1949; 1967; Schlick, 1936; 1979, vol. 2; Waismann, 1945; Wittgenstein, 1979).

3 In the *Tractatus* (Wittgenstein, 1961, 6. 211) he merely notes, "In philosophy the question, 'What do we actually use this word or this proposition for?' repeatedly leads to valuable insights."

The characterization of a word's meaning as its use in a language becomes prominent in his conversations with Schlick and others between 1929 and 1932 and in lectures and writing of the 1930s. For example, in *Philosophical Grammar* (Wittgenstein, 1974): "We ask 'How do you use the word, what do you do with it' – that will tell us how you understand it" (p. 87); "The use of a word in the language is its meaning … Grammar describes the use of words in the language" (p. 60). Further, description of use is description of *rules* of use, like description of 'rules of a game': "'I can use the word *yellow*' is like 'I know how to move the king in chess'" (p. 49). Wittgenstein retains this conception even when he drops his verificationism (on which see §2.1).

4 '[T]he meaning of a word is what the explanation of its meaning explains … "What 1 c.c. of water weighs is called '1 gram' – Well, what *does* it weigh?" … Meaning, in our sense, is embodied in the explanation of meaning' (Wittgenstein, 1974, pp. 59–60).

5 You need to know this basic principle to lie (to seek to make someone believe something is true which you know to be false, by asserting it).

6 "It is what is regarded as the justification of an assertion that constitutes the sense of the assertion" (Wittgenstein, 1974, p. 81).

7 Here I use 'semantics' broadly, as equivalent to 'theory of meaning' (or if the term 'meaning' is resisted, of 'language use'). In this broad sense it is not distinguished from syntax but includes it.

8 I borrow the name from Michael Dummett (e.g., Dummett, 1993c, ch. 2), but the account of the idea in what follows is my own.

9 EM should also be distinguished from conceptual role semantics (see Field, 1977, and Peacocke, 1981 and ed. 1993, for a selection of representative articles). They often sound similar, and similar issues, for example about reference and truth, arise for them. The difference is that EM describes understanding in terms of grasp of rules, while conceptual role semantics describes it solely in terms of assertoric and inferential dispositions. The difference disappears if grasp of rules reduces to assertoric and inferential dispositions – whether or not it does is *one* of the issues at stake in the rule-following considerations (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING).

10 What is abolished is the idea of epistemology as the study of norms of belief, understood as distinct from linguistic conventions or proposals. 'Epistemology' can still remain as the name for conceptual analysis of what kind of fact is asserted to hold when one says, for example, that a person *knows* that so-and-so is the case.

11 Warrant, incidentally, comes in degrees. That is important, and a comprehensive theory would need to take it into account. However, it will not be considered here.

12 It may mention them in giving truth-conditions for sentences which themselves mention concepts and propositions. But while this might show that they enter its ontology, it would not show that it makes explanatory appeal to them in exhibiting what language-understanding is.

13 This attempt to show that truth-conditional semantics is consistent with the priority thesis is loosely based on earlier discussions by Davidson and others of the philosophy underlying his program for semantics. See Davidson (1984); Evans and McDowell (1976), 'Introduction,' and Davidson's reply to Foster therein.

14 This line of thought implies either (a) a 'deflationary,' 'redundancy,' or 'minimalist' theory of truth (a theory of the kind discussed, e.g., in Horwich, 1990) or (b) a verificationist theory of truth (see §2.1). So the disjunction of (a) and (b) *follows* from the priority thesis. A separate issue is whether a truth-conditional theory which clear-headedly *rejects* the priority thesis has to adopt some theory of truth more robust than (a) (including (b) among these more robust theories). Some argue that it does have to do so (see Peacocke, 1993b, xvi); I do not myself think that is so.

15 Remember that we are talking here of a *normative* response: there is no attempt in this account of understanding to reduce or eliminate normative attitudes to language use.

16 Contemporary philosophers who have influentially espoused a view of truth like this include Putnam (1990) and Wright; but Wright now accepts it only for some areas of discourse – see his concept of 'superassertibility' in Wright (1992). Interestingly, it is not prominent in either Wittgenstein or the Vienna Circle. Schlick is closest to it. Neurath inclined to coherentism or to questioning the very respectability of the concept of truth; Carnap (1949), relying on Tarski's semantic characterization of truth, defended truth as a respectable concept but explicitly distinguished it from assertibility. Wittgenstein inclined to a deflationary view of it. As remarked in n. 14, both the verificationist and the deflationary view are consistent with the priority thesis.

17 This passage comes from a section copied by Stein from notes which Waismann circulated as a transcript of Wittgenstein's views (see editor's introduction to Wittgenstein, 1979, p. 20).

18 "The direction of a thought-movement is defined by the logical place of the answer." Note the continuity with *Tractatus* 6.5 and 6.51; cp. Wittgenstein (1975, pp. 66 and 174): 'The meaning of a question is the method of answering it,' and 'Every proposition is the signpost for a verification.'

19 In Waismann's own theses (included in Wittgenstein, 1979, as Appendix B) the 'positivistic' kind of verificationism is rather more prominent – "To understand a proposition means to know how things stand if the proposition is true" – but the operationist conception is simultaneously stressed.

20 The suggestion is canvassed by Crispin Wright in "Anti-realism and revisionism," *Realism, Meaning and Truth*, pp. 317–341. Cp. Skorupski (1988, §VI, pp. 516–523).

21 Further discussion of related issues, together with further reading, can be found in Wright (1987, pp. 309–316: "Could Thatcher be a master-criminal?"). Note, however, that Wright's discussion is about the implications for an 'epistemically constrained' notion of truth, whereas here the issue concerns EM and its account of the meaning of logical operators.

22 In this section I have skirted obscure and much-discussed issues about 'holism,' 'anti-realism,' and classical logic. A statement of Dummett's view (which is opposed to that taken here) is Dummett (1991, chs 8–12). Compare Wright, "Anti-realism," §VI in Wright (1987). The line I have taken is discussed a little more extensively in §§I–VII of Skorupski (1993a).

23 "What else but a fact can a statement express? In what sense could something be called 'true' or 'false' if it does not designate an existing or nonexisting fact?" (Carnap, 1967, p. 341).

24 It does not correspond to the fact that the patch is yellow. Rather, the English sentence 'The patch is yellow' can express a fact because the normative proposition 'If you're following the rules of English you ought to call this patch "yellow"' can be determinately true.

25 Two writers who argue for an epistemic theory of content – though in quite different and, indeed, unrelated ways – are John Pollock and Christopher Peacocke. Peacocke's theory has been developed in a number of writings, most recently at book length in Peacocke (1992). An accessible account of Pollock's view is in Pollock (1987). Also, those writers in the Davidsonian truth-conditional tradition who hold that interpreting the meaning of a speaker's utterances requires that one attribute norms of rationality to the speaker, in effect yoke a truth-conditional semantics to an epistemic theory of content. See Davidson (1984).

26 For further discussion of these matters see Coffa (1991) and Skorupski (1993b). See also Chapter 23, ANALYTICITY.

27 The dispute goes right back to the Vienna Circle. Neurath announced that "*Statements are compared with statements*, not with 'experiences,' 'the world,' or anything else" (Neurath, 1959, p. 291). Schlick replied: "It is my humble opinion that we can compare anything to anything if we choose" (Schlick, 1979, vol. 2, p. 401). See Jacob (1984). A non-linguistic version of the Neurathian doctrine is that only a belief can provide a reason for a belief: see, e.g., Davidson (1986). For recent discussions of how experience provides reasons for belief see McDowell (1994) and Millar (1991).

28 Peacocke argues that a theory of concepts cast in terms of norms of reasoning can yield a substantive account of the *a priori* which differs from the EM account of analyticity discussed in this section. His account does not seem to me to be successful, but neither does it seem to me to be needed for his project of stating possession conditions for concepts. See Peacocke (1993a), Skorupski (1995), and Peacocke (1996).

References

Baker, G. P. 1974. "Criteria: a new foundation for semantics." *Ratio*, 16: 156–189.

Baker, G. P., and P. M. S. Hacker. 1980. *Wittgenstein, Understanding and Meaning*. Oxford: Blackwell.

Blackburn, S. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Oxford University Press.

Carnap, R. 1936 and 1937. "Testability and meaning." *Philosophy of Science*, 3: 419–471; 4: 1–40.

Carnap, R. 1949. "Truth and confirmation." In Feigl and Sellars, 1949, pp. 119–127. Translation, with adaptations, of "Wahrheit und Bewahrung," in *Actes du Congnès International de Philosophie Scientifique* (1936).

Carnap, R. 1967. "Pseudo-problems in philosophy: the heteropsychological and the realism controversy." In *The Logical Structure of the World*, translated by Rolf A. George. London: Routledge and Kegan Paul.

Coffa, A. 1991. *The Semantic Tradition from Kant to Carnap: To the Vienna Station*. Cambridge: Cambridge University Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, D. 1986. "A coherence theory of truth and knowledge." In *Truth and Interpretation*, edited by E. Lepore, pp. 307–319. Oxford: Blackwell.

Dummett, M. 1959a. "Truth." *Proceedings of the Aristotelian Society*, 59: 141–162. Reprinted in Dummett 1978a.

Dummett, M. 1974. "What is a theory of meaning?" In *Mind & Language*, edited by Samuel Guttenplan. Oxford: Oxford University Press. Reprinted in Dummett, 1993a, as "What is a theory of meaning? (I)" with an appendix.

Dummett, M. 1978a. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. 1978b. "The philosophical basis of intuitionistic logic." In Dummett, 1978a, pp. 215–247.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. London: Duckworth.

Dummett, M. 1993a. *The Seas of Language*. Oxford: Oxford University Press.

Dummett, M. 1993b. "Realism and anti-realism." In Dummett, 1993a, pp. 462–478.

Dummett, M. 1993c. *Origins of Analytical Philosophy*. London: Duckworth.

Dummett, M. 1994. "Wittgenstein on necessity: some reflections." In *Reading Putnam*, edited by Bob Hale and Peter Clark, pp. 49–65. Oxford: Oxford University Press. Also in Dummett, 1993a, pp. 446–461.

Evans, G., and J. McDowell, eds. 1976. *Truth and Meaning: Essays in Semantics*. Oxford: Clarendon Press.

Feigl. H., and W. Sellars, eds. 1949. *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts.

Field, H. 1977. "Logic, meaning and conceptual role." *Journal of Philosophy*, 74(34): 7–75.

Hacker, P. M. S. 1990. *Wittgenstein, Meaning and Mind: An Analytical Commentary on the Philosophical Investigations*, vol. 3. Oxford: Blackwell.

Horwich, P. 1990. *Truth*. Oxford: Blackwell.

Horwich, P. 1995. "Meaning, use and truth." *Mind*, 104(414): 355–368.

Jacob, P. 1984. "The Neurath–Schlick controversy." *Fundamenta Scientiae*, 5: 351–366.

McDowell, J. 1994. *Mind and World*. Cambridge, MA: Harvard University Press.

Millar, A. 1991. *Reasons and Experience*. Oxford: Oxford University Press.

Moore, G. E. 1959. "Wittgenstein's lectures in 1930–33." In *Philosophical Papers*. London: George Allen and Unwin.

Neurath, O. 1959. "Sociology and physicalism." In *Logical Positivism*, edited by A. J. Ayer, pp. 282–317. Glencoe, IL: Free Press; London: Allen and Unwin.

Peacocke, C. 1981. "The theory of meaning in analytic philosophy." In *Contemporary Philosophy, vol. 1, Philosophy of Language, Philosophical Logic*, edited by G. Floistad, pp. 57–82. The Hague: Martinus Nijhoff.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. 1993a. "How are a priori truths possible?" *European Journal of Philosophy*, 1(2): 175–199.

Peacocke, C. ed. 1993b. *Understanding and Sense*, vol. 1. Aldershot: Dartmouth.

Peacocke, C. 1996. "Can a theory of concepts explain the *a priori*? A reply to John Skorupski's critical notice of *A Study of Concepts*." *International Journal of Philosophical Studies*, 4(1): 154–160.

Pollock, J. 1987. *Contemporary Theories of Knowledge*. London: Hutchinson.

Putnam, H. 1983. "Computational psychology and interpretation theory." In *Realism and Reason: Philosophical Papers*, vol. 3, pp. 139–154. Cambridge: Cambridge University Press.

Putnam, H. 1990. *Realism with a Human Face*. Cambridge, MA: Harvard University Press.

Schlick. M. 1936. "Meaning and verification." *The Philosophical Review*, 45(4): 339–369. Reprinted in Feigl and Sellars, 1949, and Schlick, 1979.

Schlick. M. 1979. *Philosophical Papers* (2 vols), edited by Henk L. Mulder and Barbara F. B. van de Velde-Schlick, translated by Peter Heath. Dordrecht, Netherlands: Reidel.

Skorupski. J. 1988. "Realism, meaning and truth" (critical review of Wright, *Realism, Meaning and Truth*). *Philosophical Quarterly*, 38(53): 500–525.

Skorupski. J. 1993a. "Anti-realism, inference and the logical constants." In *Realism and Reason*, edited by J. Haldane and C. Wright, pp. 133–164. Oxford: Oxford University Press.
Skorupski. J. 1993b. *English-Language Philosophy 1750–1945*. Oxford: Oxford University Press.
Skorupski. J. 1995. "Possessed by concepts" (critical notice of Christopher Peacocke, *A Study* of *Concepts*). In *International Journal of Philosophical Studies*, 3(14): 3–64.
Skorupski. J. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.
Strawson. P. F. 1977. "Scruton and Wright on anti-realism, etc." *Proceedings of the Aristotelian Society*, 77(1): 15–22.
Stroud, B. 1965. "Wittgenstein and logical necessity." *Philosophical Review*, 74: 504–518.
Waismann, F. 1945. "Verifiability." *Proceedings of the Aristotelian Society*, suppl. vol. 19: 119–150.
Wittgenstein, L. 1958. *Philosophical Investigations*, 2nd edn, edited by G. E. M. Anscombe and R. Rhees, translated by G. E. M. Anscombe. Oxford: Blackwell.
Wittgenstein, L. 1961. *Tractatus Logico-Philosophicus*, translated by D. P. Pears and B. F. McGuinness. London: Routledge and Kegan Paul.
Wittgenstein, L. 1974. *Philosophical Grammar*, edited by Rush Rhees, translated by Anthony Kenny. Oxford: Blackwell.
Wittgenstein, L. 1975. *Philosophical Remarks*, edited by Rush Rhees, translated by Raymond Hargreaves and Roger White. Oxford: Blackwell.
Wittgenstein, L. 1978. *Remarks on the Foundations of Mathematics*, 3rd edn, edited by G. H. von Wright, R. Rhees, and G. E. M. Anscombe, translated by G. E. M. Anscombe. Oxford: Blackwell.
Wittgenstein, L. 1979. *Wittgenstein and the Vienna Circle,* conversations recorded by Friedrich Waismann, edited by Brian McGuiness, translated by Joachim Schulte and Brian McGuiness. Oxford: Blackwell.
Wright. C. 1987. *Realism, Meaning and Truth*. Oxford: Blackwell.
Wright. C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Postscript

BERNHARD WEISS

John Skorupski's chapter leads us from broad ideas about meaning as being constituted by use to the verificationist conception of meaning (§2). The verificationist account takes it that we can characterize the meaning of a sentence in terms of the conditions warranting its assertion. However, as Skorupski notes, this conception seems to face insurmountable obstacles. And thus we are led 'beyond verificationism' (§2.3). Here I reprise these difficulties for verificationism and look at recent attempts to respond to them within the 'meaning as use' paradigm. I end with a brief review of recent discussion of Fitch's paradox, which appears to challenge, what is often taken to be a consequence of a use-conception of meaning, verificationism about truth.

Meaning Is Use

It's very easy to agree with the slogan "meaning is use" simply because words don't select their own meanings. Words, as shapes or sounds, are intrinsically meaningless and acquire meanings on being put to a multitude of uses by speakers; we, speakers, confer meanings on our words by *using* them, in context, in the ways that we do. And differences in the use of words go along with differences in their meaning. But historically the slogan has been espoused with more positive intent than just this.

The slogan can acquire more substance either by linking it to a methodological approach or to explanatory ambitions. Methodologically, we might find ourselves being told to clarify the meaning of a word by attending to how it is used; and, since the meaning of a word is a concept, we might clarify concepts by attending to the use of words. Philosophical problems often (perhaps always) require a reflective appreciation of certain elusive concepts. We can thus partially approach those problems by focusing on the use of certain vital words which express them. Moreover, one might argue that philosophical problems *only* require reflective appreciation of the nature of concepts in order to be solved or dissolved. So, one might claim, philosophical problems are dealt with by examining the use of words. Claims of this ilk are distinctive of ordinary language philosophy, and, on some readings of the later Wittgenstein, form at least part of his conception of philosophical method. I won't be discussing these approaches here; rather I want to consider meaning as use when considered in connection with explanatory ambitions.

One might accept the close connection between use and meaning but treat it simply as a *constraint* on an account of meaning. In other words, we could simply treat the observation as a way of testing the adequacy of any explanation of meaning: a difference in use should go along with a difference in the explanation of meaning. But the explanation of meaning might focus on properties of signs which differ from use, such as their relations to non-linguistic entities. Accounts of this form might be termed truth-conditional or representational theories of meaning and are described elsewhere in this volume (see Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S, and Chapter 4, MEANING, USE, VERIFICATION, §1.4). Though I'll return to these accounts later, let us now simply note that accounts which take the semantic concepts as explanatorily basic count as rejections of the *meaning as use* slogan, since meaning is explained in terms of properties other than use.

In contrast, to adhere to the meaning as use slogan, construed in the explanatory vein, is to attempt to account for meaning in terms of properties of use. Taking it that meaning *is* use, the aim of the theorist of meaning is systematically to characterize the use of all expressions in a natural language.

Use and Assertion

Words have only one kind of use: they are employed in sentences. But sentences come in a range of moods which are used in performing a range of speech acts. In one way or another meaning as use theorists privilege the role of assertion. Historically, this may be a hangover from the truth-conditional approach or may be a product of the fact that some use-theoretic accounts are attempts to generalize the approach from mathematics to the rest of language. These historical reasons are not nugatory. The reasons for focusing on truth are often good reasons for focusing on assertion, and, whatever account is settled on for language as a whole, will have to apply to the special region of mathematical language too. So, for instance, it's hard to see how commands could occupy the same role as do assertions, since it is hard to see some contents as being fit to be commanded – content about mathematical matter, but others too – and hard to see how to account for the fulfilment conditions of commands except in terms of the truth-conditions (or meaning) of corresponding assertions.

So (as Skorupski notes (§1.3)) central to use-theoretic accounts is the attempt to explain assertion conditions. As I've hinted, the approach has a provenance in the philosophy of mathematics, where the aim is to characterize the proof-conditions of a complex sentence

in terms of those of its components. It's worth pausing briefly to consider the mathematical case, if only to salvage a sense of the problems that attend the attempt to generalize it. So, for example, intuitionists in the philosophy of mathematics put forward clauses in the following style:

A proof of P ∨ Q is a construction which recognizably yields a proof of P or a proof of Q.

The clause explains what it is for a construction to be a proof of a complex sentence in terms of what it is to be a proof of its components. Anyone who understands the full gamut of clauses is able to recognize whether or not a presented construction is or is not a proof of a given sentence. And, knowing the epistemological significance of *proof*, this feeds directly into her use of the sentence in terms of whether or not she is prepared to assert it. Truth plays no explanatory role in the account; so we might be deflationists about truth but we might instead identify truth with provability: a sentence is true iff it is provable. This question will hinge on other commitments.

The reason we can identify truth with provability is that proof is indefeasible in the following sense. Take the two sentences P and 'It is provable that P,' and suppose each is assertible. Imagine now that we learn something which we legitimately take to undermine the assertibility of the one. Then the assertibility of the other is likewise undermined. This is a major difference between proof in mathematics and any plausible notion of warranted assertibility applicable to empirical discourse.

This fact generates the following problem.[1] Take the supposition that the meaning (or the meaning-determining uses) of a sentence are captured by stating its conditions of warranted assertion.

It is clear then that the sentences P and 'It is warrantedly assertible that P' share the same conditions of warranted assertibility: if one is warrantedly assertible so is the other. But the two sentences embed differently in complex sentences and thus have different meanings. To take an example, the following sentence is almost certainly true:

If there will be no rain this season then the crops will fail.

Yet the following sentence may well not be true:

If it is warrantedly assertible that there will be no rain this season then the crops will fail.

The reason we might have these different reactions to the conditional is that there is a difference between the statements expressed by (simultaneous) utterances of 'There will be no rain this season' and 'It is warrantedly assertible that there will be no rain this season.' The former may turn out to be false when it turns rainy, yet the latter remain true: the epistemic situation, as it then was, did indeed warrant the assertion; and contrariwise the former may be true when it stays dry, yet there is no good warrant for the assertion. This doesn't, however, solve the problem; it merely gives us a description of it, since talking about the statement expressed by an assertion presupposes an appreciation of its meaning. What we need is to latch onto conditions of use which enable us to mark this distinction in meaning.

There are two broad approaches to the problem. The first is to try to find a use-condition which mimics the indefeasibility of mathematical proof; the other is to complicate the set of use-conditions that we invoke in explaining meaning. Let's look at each approach in turn.

A First Response: Specialized Assertion Conditions

The first approach might be pursued by idealizing one's epistemic situation. So Peirce suggests that we think, not of what is currently assertible at our imperfect stage of enquiry, but of what is assertible at the end of enquiry. At that stage, by hypothesis, we contemplate no further amplification of enquirers' body of information and thus defeat of a statement assertible at the end of enquiry is impossible. And Putnam (1981) suggests that we focus on the ideal justification of a statement. Both of these proposals are problematic: Peirce's because it is far from evident that there is a single stage of enquiry which will simultaneously be epistemically perfect for every statement; and Putnam's because of difficulties in articulating what would be an ideal justification. In addition, both proposals appear to be susceptible to a version of the conditional fallacy. On the Peircian account we might have "P iff (if C were to obtain then it would be assertible that P)," where C are ideal conditions. But then letting P be 'C will never obtain' we arrive at: "C will never obtain iff (if C were to obtain then it would be assertible that C will never obtain)," which is patently unacceptable. Putnam seems to be in a better position because his ideal conditions, C, vary from one proposition to another. But this only helps if it can be guaranteed that for all propositions, P, the ideal conditions for judging that P obtains differ from the ideal conditions for judging that P's ideal conditions obtain. If not, we can form a restricted version of the very same problem (see Plantinga, 1982, and Wright, 2000). On Putnam's account we have:

> P iff (if conditions were to be C(P) then P would be assertible)

Let P = C(P)will never obtain

> C(P) will never obtain iff (if conditions were to be C(C(P))then it would be assertible that C(P) never obtain)

Suppose C(C(P)) = C(P) and we infer:

> C(P) will never obtain iff (if conditions were to be C(P) then it would be assertible that C(P) never obtain)

And this, as before, is unacceptable.

Finally, because these are ideal conditions which may not be humanly feasible to achieve, it is doubtful that they happily unpack a notion of use-conditions (see Skorupski, §2.1; Moore, 2012; and Williamson, 2006).

In response to these problems Wright suggests that we don't idealize conditions of assertion; rather we focus on actual conditions of warranted assertion but envisage the properties of the assertion to be suitably ideal. So he invents the notion of superassertibility: a statement is superassertible when it is warrantedly assertible and warrant for its assertion will, in fact, survive future investigation no matter how strenuous. Superassertible statements are, like proven statements in mathematics, enduringly warrantedly assertible. But Wright's account is not without difficulties: one is that *types* of sentences don't have superassertibility conditions, since the same type of condition may be a superassertion condition in one circumstance – since *in fact* it cannot be defeated on further investigation – but not in another – since in these circumstances it is *in fact* capable of defeat. But surely, in understanding a language, what speakers learn is the meaning of types of sentences[2] and

thus what the theory of meaning should aim, primarily, to detail are the meanings of types of sentences. Superassertion conditions seem ill-fitted to this role.

Before moving to the other broad category of approaches it is worth observing that one perhaps ought to read the above accounts as offering analyses of truth in epistemological terms and then employing the epistemic conception of truth in a truth-conditional account of meaning. Indeed, Peirce advances his account as a view of truth, as does Putnam; and Wright not only suggests that we view superassertibility as an epistemic conception of truth but sketches a truth-conditional theory of meaning in Davidsonian style taking this as the operative notion of truth (Wright, 1993, essay 14). So here we bring together representationalism about meaning with meaning as use, through verificationism about truth.[3]

A Second Response: Adding to Assertion Conditions

The second broad set of approaches complicates the set of uses which are taken to determine meaning. Though there are a number of flavors to choose from, all the options look beyond mere assertion conditions to another set of use-conditions. Some proposals appeal to two well-defined sets of conditions; for example some argue in favor of using both assertion and denial conditions[4]; others suggest assertion and retraction conditions (Weiss, 2007; Edgington, 1981); and Brandom situates meaning in a complex normative practice whose basic normative statuses are the commitments and entitlements which accrue as a result of making a move in the practice (see Brandom, 1994; 2000). Others simply choose a more indefinitely specified set of use-conditions. Cozzo and Horwich[5] both focus on characterizing the meaning of words through aspects of their use, Cozzo invoking the word's 'immediate argumental role' (Cozzo, 1994) and Horwich the word's acceptance conditions.[6] In either case we consider a range of such use-conditions which suffice to explain the word's overall use; and, of course, it is an assumption of the approaches that some such set of uses will indeed be explanatorily sufficient.[7]

On all these approaches meanings involve marrying sets of conditions, for instance, marrying assertion with denial conditions. And there is a degree of freedom in how meanings bring about this marriage, since, were we able to derive the one set of conditions from the other, these would not be distinct. Some treat this as a promising feature of the approach, others as problematic. Dummett[8] has long argued against what we might call multilateral[9] approaches precisely on these grounds because he thinks that we might then have a situation in which an assertion is neither correct – since its assertion conditions are not fulfilled – nor incorrect – since its denial conditions are not fulfilled; yet it is absurd to suppose that the act of assertion might have this kind of indeterminate outcome.[10] But even Dummett concedes that, if we are to account for language as it is actually used, we may need to adopt a multilateral approach (see Dummett, 2010, p. 226). Thus it seems the basis of his view is a sense that any such language is logically objectionable because it incorporates substantial presuppositions about the world. And indeed, most advocates of multilateralism accept this,[11] but tend to treat it is an advantage of use-conditional accounts over truth-conditional accounts.[12]

Fitch's Paradox or the Paradox of Knowability

I turn now to truth. The view that truth amounts to some conception of verifiability is challenged by Fitch's paradox (Fitch, 1961). Fitch's reasoning seems to show that the following is a theorem: $(\forall p)(p \rightarrow \Diamond Kp) \rightarrow (\forall p)(p \rightarrow Kp)$. And this appears to be absurd: from the claim

that any truth is knowable – by no means evidently absurd – we infer the absurd claim that every truth is known. This is certainly a problem for anyone advocating the claim that every truth is knowable but is also a problem for anyone who thinks, as most do, that knowability falls short of being known (see Brogaard and Salerno, 2007). In essence, the paradox asks us to rescue from obvious falsehood the philosophical position that truth coincides with verifiability.

I won't present the reasoning formally here. The crux of the argument is to take the proposition $q \wedge \neg Kq$ and to assume it to be true. If every truth is knowable it follows that $\Diamond K(q \wedge \neg Kq)$. But using the seemingly uncontroversial principles that knowledge is factive and that if a conjunction is known then each conjunct is known (together with weak modal principles: necessitation and closure of possibility under strict implication) we then infer that a contradiction $Kq \wedge \neg Kq$ is possible, which is, of course absurd.

Our options are as follows:

1. To reject the modal and epistemic principles involved.
2. To reject the methods of reasoning involved.
3. To reinterpret the original claim: $(\forall p)(p \rightarrow \Diamond Kp)$.

Remarks on each option:

1. The problem with rejecting the modal principles is that they are very weak. Necessitation and the closure of possibility under strict implication are principles of the weakest of modal systems, K. The epistemic principles seem equally undeniable; the incompatibility of knowing a claim with that claim's falsity is highly plausible and (given classical logic) entails that knowledge is factive. And it is hard to see how one could know a conjunction without knowing each conjunct.
2. One might reject the reasoning involved by, say, eschewing classical in favor of intuitionistic logic. But intuitionistically we would be able to infer $\neg Kp \rightarrow \neg p$, which seems equally problematic.
3. We might reinterpret the original claim. Reinterpretative strategies come in the following forms:
   a. Restriction of the quantifier: (e.g., Dummett, 2001; Tennant, 2002).
   b. Reinterpretation of the conditional and/or negation: (e.g., Dummett, 2007; de Vidi and Solomon, 2001; Williamson, 1982).
   c. Reinterpretation of the '$\Diamond$': e.g., Pragmatic versus Semantic impossibilities (e.g., Hand, 2009).
   d. Reinterpretation of 'K': e.g., Conjunctive Knowability (Restall, 2009); Recognizing the truth-conferring state of affairs (Jenkins, 2009); Non-actual knowers (Edgington, 1985).

   Reinterpretation is obviously only of service if it can be shown where, on the reinterpretation, the Fitch reasoning fails and that, once reinterpreted, the knowability principle can be squared with a plausible understanding of the claim that every truth is knowable.

Notes

1 See Brandom (1976); Dummett (1981), where he contrasts content with ingredient sense; and Wright, 1993. Skorupski points out the problem at §2.3. For a more recent debate see Casalegno (2002), Wright (2012), and Williamson (2012a; 2012b). Dummett (2002a) draws attention to the phenomenon and, interestingly, uses it *against* truth-conditional accounts of meaning.

2 Or better, what they learn is the meaning of words, which yields grasp of the meanings of types of sentences.
3 See discussion of Fitch's paradox below.
4 Rumfitt (2000) and Price (1983). For response see Dummett (2002b).
5 While most use-conditional theorists conceive of use in normative terms, interpreting conditions of use as akin to *rules* for use, Horwich is an exception and derives his acceptance properties of words from regularities of their use.
6 Horwich (1998). In this vein see also Peregrin (1995; 2006).
7 The approach can be seen to be a generalization of proof-theoretic or inferentialist approaches to the meanings of the logical connectives. Here the debate has an additional dimension in accounting not only for the meaning of the logical constants but also for the epistemology of inference. See: Boghossian (2003) and (2012); Williamson (2003; 2012a); Casalegno (2004); Peregrin (2010); Brandom (2008); Read (2004); Dummett (1991); Restall (2005), Tennant (2005).
8 A good example is his (2002a) response to Rumfitt.
9 Following Rumfitt (2000).
10 See Rumfitt (2000) and Price (1983) for rebuttals of this objection. It is also not evident that other approaches, such as marrying assertion with defeating conditions, lead to this problem.
11 Cozzo thus speaks about the "correctness of a language"; Brandom speaks about inferences which are incorporated in moving from an assertion's grounds to its consequences, and notes that these ought to be actually materially good but need not be formally good; Horwich notes the existential presupposition in some sets of acceptance conditions.
12 Terms which are, for instance, pejorative might have meanings which certain speakers find objectionable. The orthodox truth-conditional account can allow for this by relegating this to a feature of meaning irrelevant to semantic evaluation, such as Fregean color or tone; but the phenomenon discussed in the text cannot be relegated to the non-semantic in this way.

References

Boghossian, P. 2003. "Blind reasoning." *Proceedings of the Aristotelian Society*, suppl. 77: 225–248.
Boghossian, P. 2012. "Inferentialism and the epistemology of logic: reflections on Casalegno and Williamson." *Dialectica*, 66(2): 221–236.
Brandom, R. 1976. "Truth and assertibility." *The Journal of Philosophy*, 83: 137–149.
Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.
Brandom, R. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.
Brandom, R. 2008. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: Oxford University Press.
Brogaard, B., and J. Salerno. 2007. "Knowability, possibility and paradox." In *New Waves in Epistemology*, edited by D. Pritchard and V. Hendricks, pp. 270–299. London: Palgrave Macmillan.
Casalegno, P. 2002. "The problem of non-conclusiveness." *Topoi*, 21(1–2): 75–86.
Casalegno, P. 2004. "Logical concepts and logical inferences." *Dialectica*, 58(3): 395–411.
Cozzo, C. 1994. *Meaning and Argument*. Stockholm: Almqvist and Wiksell.
De Vidi, D., and G. Solomon. 2001. "Knowability and intuitionistic logic." *Philosophia*, 28(1–4): 319–334.
Dummett, M. 1981. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.
Dummett, M. 1991. *The Logical Basis of Metaphysics*. London: Duckworth.
Dummett, M. 2001. "Victor's error." *Analysis*, 61(269): 1–2.
Dummett, M. 2002a. "Meaning in terms of justification." *Topoi*, 21(1): 11–19.
Dummett, M. 2002b. "'Yes,' 'no' and 'can't say.'" *Mind*, 111(442): 289–295.
Dummett, M. 2007. "Reply to Wolfgang Künne." In *The Philosophy of Michael Dummett*, edited by R. E. Auxier and L. E. Hahn, pp. 345–350. Chicago: Open Court.

Dummett, M. 2010. "Should semantics be deflated?" In *Reading Brandom: On Making it Explicit*, edited by B. Weiss and J. Wanderer, pp. 213–226. London: Routledge.

Edgington, D. 1981. "Meaning, bivalence and realism." In *Proceedings of the Aristotelian Society*, 81: 153–173.

Edgington, D. 1985. "The paradox of knowability." *Mind*, 94(376): 557–568.

Fitch, F. 1961. "A logical analysis of some value concepts." Retiring presidential address presented to the Association for Symbolic Logic, Atlantic City, NJ, December 23, 1963. Frederic B. Fitch Papers, Box 33, Manuscripts and Archives, Yale University Library, New Haven, CT.

Hand, M. 2009. "Performance and paradox." In Salerno 2009, pp. 283–301.

Horwich, 1998. *Meaning*. Oxford: Clarendon Press.

Jenkins, C. 2009. "The mystery of the disappearing diamond." In Salerno, 2009, pp. 302–319.

Moore, A. 2012. "Dummett: the logical basis of metaphysics." In *The Evolution of Modern Metaphysics: Making Sense of Things*, pp. 345–368. Cambridge: Cambridge University Press.

Peregrin, J. 1995. *Doing Worlds with Words*. Dordrecht, Netherlands: Springer.

Peregrin, J. 2006. "Meaning as an inferential role." *Erkenntnis*, 64(1): 1–36.

Peregrin, J. 2010. "Inferentializing semantics." *Journal of Philosophical Logic*, 39(3): 255–724.

Plantinga, A. 1982. "How to be an anti-realist." *Proceedings and Addresses of the American Philosophical Association*, 56(1): 47–70.

Price, H. 1983. "Sense, assertion, Dummett and denial." *Mind*, 92(366): 174–188.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Read, S. 2004. "Identity and harmony." *Analysis*, 64(2): 113–119.

Restall, G. 2005. "Multiple conclusions." In *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress*, edited by P. Hájek, L. Valdés-Villanueva, and D. Westerståhl, pp. 189–205. London: King's College Publications.

Restall, G. 2009. "Not every truth can be known (as least, not all at once)." In Salerno, 2009, pp. 339–354.

Rumfitt, I. 2000. "'Yes' and 'no.'" *Mind*, 109(436): 781–823.

Salerno, J., ed. 2009. *New Essays on the Knowability Paradox*. Oxford: Oxford University Press.

Tennant, N. 2002. "Victor vanquished." *Analysis*, 62(274): 135–142.

Tennant, N. 2005. "Rule circularity and the justification of deduction." *The Philosophical Quarterly*, 55(221): 625–648.

Weiss, B. 2007. 'Anti-realist truth and anti-realist meaning.' *American Philosophical Quarterly*, 44(3): 213–228.

Williamson, T. 1982. "Intuitionism disproved?" *Analysis*, 42(4): 203–207.

Williamson, T. 2003. "Understanding and inference." *Proceedings of the Aristotelian Society,* suppl. vol. 77: 249–293.

Williamson, T. 2006. "Must do better." In *Realism and Truth*, edited by P. Greenough and M. Lynch, pp. 177–187. Oxford: Clarendon Press.

Williamson, T. 2012a. "Wright and Casalegno on meaning and assertibility." *Dialectica*, 66(2): 267–271.

Williamson, T. 2012b. "Boghossian and Casalegno on understanding and inference." *Dialectica*, 66(2): 237–247.

Wright, C. 1993. *Realism, Meaning and Truth*, 2nd edn. Oxford: Blackwell.

Wright, C. 2000. "Truth as sort of epistemic: Putnam's peregrinations." *The Journal of Philosophy*, 97(6): 335–364.

Wright, C. 2012. "Meaning and assertibility: some reflections on Paolo Casalegno's 'The problem of non-conclusiveness.'" *Dialectica*, 66(2): 249–266.

5

Semantics and Pragmatics

GUY LONGWORTH[1]

1 Pragmatics and Semantics

We use language in order to achieve a wide variety of ends. To a good first approximation, pragmatics is the study of the things we do in, or by, using language, together with the facts, knowledge, and abilities that we exploit in order to do those things. Correlatively, pragmatic phenomena are all and only the phenomena that are studied in doing pragmatics. In using bits of language in order to achieve our ends, we rely on various facts about those bits of language, together with knowledge and abilities relating to those facts. For one prominent example, we exploit facts about the meanings of the words that we use, and facts about the meanings of sentences that depend upon the combination of words in them given those words' meanings. To a good first approximation, semantics is the study of facts about the meanings of words and sentences, about the dependence of sentence meaning on combinations of word meanings, and about our knowledge and abilities relating to those facts. Correlatively, semantic phenomena are all and only those phenomena studied in doing semantics. There is no obvious reason, in advance of further inquiry, to suppose that no phenomena fall within the remits of both semantics and pragmatics, and some reason to think that many phenomena fall within both. For example, it is natural to assume that speakers' knowledge of meaning will figure in shaping their use of language, and in the reception of that use. And many philosophers have held that the use of language plays a constitutive role in determining the standing meanings of the words that are so used. However, some theorists appeal to a more restrictive conception of pragmatics – roughly, pragmatics minus semantics – and thereby to force a sort of partition. Nothing turns on this.

(The label "pragmatics" derives from the Greek root "pragma," meaning *deed*. The earliest relevant occurrence in the *Oxford English Dictionary* for the study of practical aspects of human action and thought is the 1693 title of a book by E. Settle, *The new Athenian comedy containing the politicks, oeconomicks, theologicks, poeticks, mathematicks, sophisticks, pragmaticks, dogmaticks, &c. of that most learned society*. The label "semantics" derives from the

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Greek roots "sēma," meaning *sign*, or "sēmantikos," meaning *significant*. The earliest relevant occurrence in the *Oxford English Dictionary*, for that which relates to divination through the interpretation of signs, is from 1665, in John Spencer's *A discourse concerning prodigies* (2nd edn, London: printed by F. Field for W. Graves). Contemporary uses of the labels seem to derive, more or less loosely, from work by Charles Morris in the 1930s, culminating in his (1938). But we shouldn't get too hung up about labels. Useful discussions of various ways of distinguishing between semantics and pragmatics may be found in Bach, 1999, and Szabó, 2006.)

In addition to getting on with the projects of pragmatics and semantics – that is, theorizing about things done with words and about some of the properties of words that figure in doing those things – philosophers and linguists have been concerned to say more about the natures of the two projects, the range of phenomena that fall within each, and the connections between the projects and the phenomena that each aim to treat. As Dorothy Edgington puts it,

> At its most general, the issue here is how much of our ability to communicate rests on specifically linguistic knowledge, and how large a role is played by background knowledge, common sense and inference to the best explanation, all of which play a role in the pragmatics of communication. (Edgington, 2006, p. 769)

Of central importance here have been two related questions.

The first question concerns one type of thing that people do with language and that seems to depend heavily on the meanings of the words that they use: stating or asserting things. (I'll treat stating and asserting as the same activity.) Philosophers care about what people state, and the circumstances in which they state those things, for various reasons, but in particular because of the way that the things people state figure in revealing their commitments, including their beliefs. (That is not to say that what one commits to in speaking, or in stating what one does, is to be identified with that which one says. We might say things without thereby committing ourselves, for instance in speaking ironically. And we might commit ourselves to things that we do not say or state, for instance in stating things that entail further commitments.) The first question is whether facts about what people state – for example, the fact that Bill stated that Jill smokes – are to be accounted for within semantics or pragmatics. Are the facts about what people state determined solely by the meanings of the words they use – so that they fall within the remit of semantics? Or are facts about what people state dependent also on general facts about what they are up to, the circumstances in which they speak, and so forth – so that they fall partly within the remit of pragmatics?

The second question concerns a specific mode of evaluation or description of people's actions, or of the commitments embodied in those actions: evaluation or description as to truth. We often evaluate or describe what people state, as well as what they believe, as being true or false. For instance, Bill stated that Jill smokes. But Jill doesn't smoke. So what Bill stated is false. Philosophers care about this mode of evaluation in part because it figures in wider assessments of what people do: for one artificial example, if we suppose that it is wrong to state things that are false, then the fact that what Bill stated is false entails that it was wrong of Bill to state it. In addition, philosophers care about the conditions in which one or another evaluation of that sort is mandated, the conditions in which what someone is committed to would be true or false. Bill stated that Jill smokes. How would the world need to be in order for what Bill stated to be true? How would it need to be for what he stated to be false?

In learning answers to those questions, we might learn something about what it is for someone to smoke. And so, *mutatis mutandis*, for topics of more pressing philosophical interest: the conditions in which stating that Bill *knows* that Jill smokes would be stating something true or false; the conditions in which stating that it is *wrong* for Jill to smoke would be stating something true or false; and so forth. The second question is whether facts about the conditions in which what someone states would be true or false are to be accounted for within semantics or pragmatics. Are the facts about the conditions in which what someone states would be true or false determined solely by the meanings of the words they use – so that those facts fall within the remit of semantics? Even if they are not, it may be that the meaning of a sentence determines its truth-conditions, so that it is possible to assess sentences as true or false independently of what speakers state by the use of those sentences. Alternatively, facts about truth-conditions may be dependent also on general facts about what speakers are up to, the circumstances in which they speak, and so forth – so that those facts fall within the remit of pragmatics. Crucially, it will be possible to account for sentence meaning by appeal to truth-conditions, or by appeal to propositions that determine truth-conditions, only in so far as meaning determines truth-conditions. (See Wiggins, Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S.)

In order to discuss those issues, we will need to spend some time developing the orthodox contemporary position on these matters, which derives from the work of Paul Grice (§§3–5). And it will facilitate our understanding of Grice to start a little further back, with the types of position to which Grice was responding, and an initial response to those types of view, due to J. L. Austin, that Grice sought to develop in his own work (§§2–3). Having developed the orthodox view, we'll then consider one source of opposition to the view, due to Charles Travis, and see how it leads to an alternative view (§§6–7).

2 Austin on Locutionary, Illocutionary, and Perlocutionary Acts

Contemporary recognition of the importance of divisions amongst pragmatic and semantic phenomena has its roots in earlier recognition of the importance of pragmatic phenomena. By the middle of the twentieth century, philosophers had begun to emphasize the importance of attention to language use in theorizing about language, including attention to the things speakers achieve, or aim to achieve, by saying what they do. In itself, the emphasis on use seems admirable, and is preserved in more recent theorizing about language. However, it was common during this period for philosophers to overplay the connection between meaning and use. Gripped by the importance of attention to the use of language, some philosophers went so far as to claim that meaning is determined by, and fully explains, use – in its most plausible formulations, that the meanings of many words are determined by the uses of sentences involving those words. More generally, a number of philosophers had assumed that appeals to meaning would underwrite direct explanations of use, so that it would be comparatively straightforward to discern features of the meanings of words from features of the uses of those words. (Such an assumption arguably figures in work of this period by R. M. Hare, Norman Malcolm, G. E. Moore, Gilbert Ryle, P. F. Strawson, and Ludwig Wittgenstein, amongst others.) Attempts to distinguish meaning from use, and to draw distinctions amongst types of uses of language, emerged in reaction to such assimilations of meaning and use. (See Soames, 2003, pp. 65–219.)

While it is surely true that the meanings of words play an explanatory role in many of the things that we do with those words, that bromide leaves open that the explanation of language use may have many other moving parts and that, as a consequence, the connection between meaning and use might be quite indirect. Moreover, it leaves open that the use of language might be many-faceted, and that meaning might play different roles – and more or less direct roles – in explaining different facets of use. Recognition of both points figured in the articulation of distinctions between semantic and pragmatic phenomena as well as distinctions amongst the pragmatic phenomena that meaning figures in explaining. (It is perhaps worth observing that many philosophers during this period had a more general tendency to assume that explanatory connections would be direct. For example, many philosophers during this period were sympathetic to the ideas that aspects of mind are directly expressed in behavior and that the cognitive content of scientific theories might be determined in a simple way by the pattern of observations taken to support those theories. So, the recognition of a need to allow for greater theoretical distance between meaning and various facets of use is of a piece with a more general cognizance of the potential indirectness of explanatory connections.)

J. L. Austin played an important role in directing attention onto some of the required distinctions. In particular, he sketched a distinction amongst three types of thing a speaker might do in using a sentence – that is, amongst three types of act a speaker might perform by speaking as they do:

> *The locutionary act*: the production of an utterance that can be classified by its phonetic, grammatical, and lexical characteristics, up to sentence meaning (the *phatic* act). It is also the performance of an act that can be classified by its *content* (the *rhetic* act) – a feature distinctively of acts of speech. If I promise *that I'll be home for dinner* and then promise *that I'll work late*, my actions are instances of two different locutionary acts: one with the content that I'll be home for dinner, and one with the content that I'll work late. (Austin, 1962b, pp. 94–98)
>
> *The illocutionary act*: an act classifiable not only by its content – as with the locutionary act – but also by its *force* (stating, warning, promising, etc.). If I *promise* that I'll be home for dinner and later *state* that I'll be home for dinner, my actions might be instances of a single locutionary act: both actions might involve the content that I'll be home for dinner. However, my actions are instances of different illocutionary acts: one has the force of a promise, while the other has the force of a statement. (Austin, 1962b, pp. 98–101)
>
> *The perlocutionary act*: an act classifiable by its "… consequential effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons …." If I warn that the ice is thin, and so perform one illocutionary act, I may thereby perform a variety of perlocutionary acts: I may *persuade* someone to avoid it, or *encourage* someone to take a risk, and so forth. (Austin, 1962b, p. 101)

Crucially, these are distinctions amongst types of things that one might do, rather than amongst individual actions. It is possible for a single action to instantiate all three types of act: a speaker's saying something might be their uttering some words with particular meanings, their inviting someone to leave, and their inducing a guest to feel unwelcome. Moreover, as in other cases of human action, the various things a speaker does will often be ordered as means to ends: the speaker might have uttered those words *in order to* produce

an invitation; and they might have produced the invitation *in order to* make their guest feel unwelcome. Thus, we can see how the meanings of the words that the speaker used, which figure most directly in the locutionary act that they performed, figure less directly in the illocutionary act or acts that they perform, and less directly still in the further, perlocutionary consequences of the performance of that illocutionary act.

Consider, for example, saying to a guest, "You should leave now." Performing that locutionary act might be a means to the performance of any of a variety of illocutionary acts. For instance, one might give an order or make a statement. And one might use the same words without either ordering or stating that the guest should leave, for example if one uttered those words as part of a larger construction: "If your train arrives at 10 p.m., you should leave now." Attempting to account for that variety just by appeal to meaning would lead to an implausibly complicated account of meaning. Similarly, performing one of those illocutionary acts – say, ordering the guest to leave – might be a means to the performance of any of a variety of perlocutionary acts: one might encourage the guest to leave, or to stay; one might thereby reveal one's inhospitality towards the guest, or one's hospitality to other guests; one might cause the guest's enmity, or induce in them a grudging respect; and so forth. Again, any attempt to account for that variety by appeal just to meaning would be bound to lead to implausible complications. Austin's tripartite distinction made clear that there is no reason to try to explain all the things that can be done by the use of some words by appeal only to the meanings of those words. Unlike locutionary acts, illocutionary and perlocutionary acts are to be explained not only by appeal to meanings, but also by appeal to a combination of further factors, including speakers' illocutionary and perlocutionary ends, and other features of the broader circumstances in which the act of speaking takes place. However, Austin's discussion left open the precise nature of the distinctions between locutionary, illocutionary, and perlocutionary acts, and provided inadequate guidance to the application of those distinctions. (See Bird, 1981; Hornsby, 1988; 1994; 2006.)

3 Grice on Illocutionary Acts

Paul Grice's work on what he called "non-natural meaning" provided the basis for one way of trying further to develop Austin's three-way distinction (Grice, 1989, pp. 213–223). As we'll see, it also furnished a motivation for drawing further distinctions within the field of illocutionary acts. Grice sought to contrast non-natural meaning with natural meaning. An example of natural meaning would be the connection between black clouds and impending rain that is specified when one claims that those black clouds mean rain. To a good first approximation, the connection is *factive* – if rain is not forthcoming, then it wasn't the case that the clouds meant rain. And except in special cases, the connection is independent of anyone's beliefs or intentions: the fact that the clouds mean rain has nothing to do with what anyone believes or intends about the clouds or about the rain. (Special cases might include cases in which someone's behavior means that they intend or believe something.) Examples of non-natural meaning would be the fact that those three rings on the bell mean the bus is full, or that, in saying, "The bus is full," the conductor meant that the bus is full. By contrast with natural meaning, non-natural meaning is non-factive and does appear to depend on speakers' intentions.

Grice proposed an account of non-natural meaning that was given by appeal to speakers' intentions to have certain cognitive effects on an audience on the basis of the audience's

recognition of some of the speakers' intentions. In one of Grice's formulations, appeal is made to a reflexive intention – an intention the content of which makes reference to itself:

> "*A* meant$_{NN}$ something by *x*" [that is, a speaker, *A*, non-naturally meant something by producing some behavior – say, an utterance – *x*] is (roughly) equivalent to "*A* intended the utterance of *x* to produce some effect in an audience by means of the recognition of this intention"; and we may add that to ask what *A* meant is to ask for a specification of the intended effect… (Grice, 1989, p. 220)

It is this type of formulation, given in terms of reflexive intentions, to which some theorists have appealed in trying to clarify the nature of illocutionary acts. On this account of illocutionary acts, what marks out illocutionary acts from other kinds of act is that an act is illocutionary just in case the intended audience's recognition of the intention with which the act is undertaken – the intention recognizably to perform one or another illocutionary act – suffices for its successful completion. (See Bach and Harnish, 1979; Hornsby, 1994; McDowell, 1980; Searle, 1969. For a useful critical summary of attempts to account for the nature of illocutionary acts, see Bird, 1981.) When illocutionary acts are understood in this way, they are distinguished clearly from perlocutionary acts. For example, one might act with the perlocutionary intention that one's action should induce an audience to leave, but one's audience might recognize one's intention without leaving and, so, without their recognition of one's intention sufficing for achievement of the perlocutionary end with which one acted. By contrast, if one acted with the illocutionary intention that one's audience recognize one's illocutionary intention – say, one's intention recognizably to order one's audience to leave – then it would suffice for success that one's audience did recognize that intention. Moreover, if we assume that locutionary acts, and the intentions with which they are undertaken, can be characterized without appeal to potential effects on an audience, the appeal to the presence or absence of reflexive intentions provides the basis for a clear distinction between locutionary and illocutionary acts.

Let's suppose that Grice's account of non-natural meaning can serve in this way to sharpen the boundaries of the class of illocutionary acts. The picture that emerges is one in which some of a speaker's actions instantiate at least three types of act: locutionary acts – the utterance of some meaningful words; illocutionary acts – the production of some behavior (for example, behavior instantiating a locutionary act) with the intention to make recognizable an intention to perform a certain type of illocutionary act; and perlocutionary acts – the production of some behavior (for example, an illocutionary act) with the intention of having effects over and above an audience's recognition of one's illocutionary intentions. Moreover, it is a picture on which the three types of act are typically related as means to ends: the speaker utters some words in order to perform an illocutionary act – in order, that is, to make recognizable their illocutionary intentions; and the speaker performs that illocutionary act in order to achieve further perlocutionary ends.

So stated, the account is deceptively simple. Room must be made for some additional complications that arise from the role of intentions in the account. For one thing, intentions figure in rational psychology, and are plausibly subject to rational constraint. In particular, one can rationally intend to Φ only in so far as one holds that it is possible for one to Φ. So, one can rationally intend to make recognizable one's illocutionary intention by performing a locutionary act in particular circumstances only in so far as one holds that it's possible that one's performance of the locutionary act in those circumstances will make recognizable

one's illocutionary intention. Furthermore, it's plausible that one can rationally intend to make recognizable one's illocutionary intentions only if it is *reasonable* for one to hold that it's possible that one's performance in those circumstances will make recognizable one's intention. Thus, if we build into Grice's account that meaning-determining intentions must be rational, the outcome will be one on which meaning-determining intentions are highly constrained by what it's reasonable to expect concerning the interaction of one's overt performance with the recognitional capacities of one's audience. In addition, we shouldn't expect that the intentions with which a speaker acts invariably will form a consistent set. Even amongst the fairly rational, cases are likely to arise in which a speaker's intentions can't all be satisfied. For instance, one might use the words, "He smokes," with the intentions, first, to speak about a man and, second, to speak about a salient individual who is not in fact a man. In such cases, judgment will be required in moving from facts about the speaker's intentions to facts about the illocutionary acts that they have performed.

Modulo those complications, the account seems to make space for a simple view about the relationship between locutionary and illocutionary acts, at least with respect to a range of central cases in which a speaker intends to exploit the meanings of their words in order to perform an illocutionary act. According to the simple view, word meanings are combined into sentence meanings in such a way as to determine that each sentence expresses a unique proposition. It is that connection between sentence and proposition that speakers exploit in order to reveal their illocutionary intention. In central cases, speakers intend to use a sentence as expressing whatever proposition is determined by its meaning in order to make recognizable their illocutionary intention to state, order, question, and so forth. In those cases, where the proposition determined by the meaning of *S* is that (or whether) *p*, and Φ is the illocutionary act type that the speaker intends to perform (e.g., stating, ordering, asking, and so forth), the speaker utters *S* in order to Φ that (or whether) *p*. According to this simple view, sentence meaning determines the propositional content of the illocutionary act, and the remaining shift from locutionary to illocutionary act is a matter merely of the determination of illocutionary force. So, the simple view fixes a backward route from illocutionary act content to meaning, and thereby puts the theorist in a position to discern facts about sentence meaning fairly directly from facts about the illocutionary acts that the sentence may be used to perform. On such a view, claims about meaning, what is stated, and truth-conditions will tend to run in step. Thus, distinctions between semantics and pragmatics on which semantics is held responsible either for meaning, or for what is stated, or for truth-conditions, will also tend to run in step. However, although the simple view is attractively simple, Grice argued that the assumption about the connection between meanings and the contents of illocutionary acts on which the simple view depends is not tenable.

4 Grice on Basic and Derivative Illocutionary Acts

In addition to sharpening Austin's distinctions between locutionary, illocutionary, and perlocutionary acts, Grice recognized the importance of drawing further distinctions within the field of illocutionary acts. In doing so, Grice was responding to philosophers who, as he understood them, endorsed something like the simple view of the relation between meaning and illocutionary acts sketched in the previous paragraph. Because they endorsed versions of the simple view, they were willing to draw conclusions about the

meanings of sentences on the basis of observations about the propositional contents of illocutionary acts that those sentences can be used to perform. Adherence to the simple view gave rise to two main difficulties. The first difficulty was that accounts of meaning based on the simple view tended to be very complex, reflecting, as Grice saw it, the fact that the connection between meaning and illocutionary act is less direct than the simple view makes it out to be. The second, related difficulty arose because the simple view led to treating as ambiguous words and sentences that might otherwise have seemed to express unitary meanings (Grice, 1989, pp. 3–21).

The basic difficulty here is that the form of words constituting a sentence can be used in order to perform illocutionary acts with a wide variety of different propositional contents, depending on variation in speakers' illocutionary intentions. To take an example connecting the first two difficulties, Peter Strawson had noticed that sentences of the form "It is true that *p*" were typically used in the performance of illocutionary acts in which it was suggested that someone had stated, or might state, that *p*. The simple view thus conditioned him to assume that that suggestion was determined by the meanings of sentences of that form (Strawson, 1949). However, where the form of words constituting that sentence are embedded in certain larger structures – for example, "If it is true that *p*, then *q*," or, "Either it is true that *p* or it is true that *q*" – the apparent correlation of word form and illocutionary suggestion lapses (Geach, 1965). The form of words can be used in different circumstances to perform illocutionary acts with different propositional contents. Thus, if one sought to explain the variety by appeal only to sentence meanings, one's account of meanings would again be required to be implausibly complicated.

Grice's response was based on the observation that very often, in performing actions that seem to exemplify a single locutionary act, speakers thereby perform a plurality of illocutionary acts. Moreover, he observed that that plurality is typically organized so that some of the basic illocutionary things speakers do serve as means to other, derivative illocutionary things that they do. Thus, to modify one of Grice's own examples, a speaker might respond to a query about the philosophical competence of one of their students by uttering the words, "They have neat handwriting." In performing that locutionary act, the speaker would be performing the following basic illocutionary act: they would be stating that the student in question has neat handwriting. However, given the circumstances in which the speaker performed that act, they might also be performing one or more derivative illocutionary acts: they might, for example, be implying, or insinuating, that the student in question was not a very good philosopher. Just as the speaker would intend the first illocutionary act to be recognizable, so they would intend the second to be. Moreover, the speaker would be performing the first illocutionary act in order to perform the second: they would intend their audience to recognize what they had stated and, on that basis, to come to recognize their further illocutionary intention to imply, or insinuate, that the student is philosophically incompetent. Thus, the new distinction within the field of illocutionary acts mirrors Austin's broader distinction between illocutionary and other acts. Grice called the propositional contents of the derivative illocutionary acts *implicatures* (Grice, 1989, pp. 3–143, 224–247; Soames, 2003, pp. 197–219).

(In addition to cases in which a speaker performs a basic illocutionary act in order to perform derivative illocutionary acts, Grice allowed for the possibility that a speaker might perform other illocutionary acts on the basis of only "making as if to" perform a basic illocutionary act (Grice, 1989, p. 30). Given the structure of Grice's account, it would be natural to treat the act of making as if to perform an illocutionary act of some type as itself

an illocutionary act, albeit of a different type. However, an alternative would be to treat making as if to perform an illocutionary act as the performance not of an illocutionary act, but only of a locutionary act. In taking that line, we would allow that some illocutionary acts are derivative not from the performance of more basic illocutionary acts, but from the performance only of locutionary acts. See Bach, 1994; 2001.)

Grice offered two sorts of accounts of cases of implicature. The first account was of what he called *conventional implicature*. Here, the central idea is that it is a feature of the meanings of certain words or sentences – or otherwise a conventional feature of those words or sentences – that their use is at the service of the performance of a plurality of speech acts (at least typically, or by default). Thus, just as a speaker can exploit the meaning of a sentence in order to help make recognizable their basic illocutionary intentions, so a speaker can exploit the fact that the use of a sentence conventionally carries certain implicatures in order to make recognizable their performance of derivative illocutionary acts with those implicatures as contents. The second account was of what he called *conversational implicature*. Grice's account of conversational implicatures is based on the idea that the ends of many central cases of linguistic interaction are, and may be presupposed by interlocutors to be, cooperative. Thus, with respect to many conversations, participants are entitled to assume that other participants aim to adhere to the following *Cooperative Principle*:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1989, p. 26)

The Cooperative Principle in turn may, in general, be implemented by appeal to a number of more specific maxims:

> *Maxims of Quantity*: 1. Make your conversational contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your contribution more informative than is required.
>
> *Maxims of Quality*: 1. Do not say what you believe to be false. 2. Do not say that for which you lack adequate evidence.
>
> *Maxim of Relevance*: Be relevant.
>
> *Maxims of Manner*: 1. Avoid obscurity. 2. Avoid ambiguity. 3. Be brief (avoid unnecessary prolixity). 4. Be orderly. (Grice, 1989, pp. 26–27)

The Cooperative Principle and its sub-maxims are used in order to help explain how speakers are able to make manifest their performance of derivative illocutionary acts on the basis of their recognizable performance of more basic illocutionary acts. (In the case of the Maxims of Manner, which make appeal to the specific ways in which speakers are to say what they do, the performance of locutionary acts also figures.) The explanation goes via the standing entitlement of speaker and audience to assume that the speaker is aiming to adhere to the Cooperative Principle and maxims. Grice assumes that the basic illocutionary act performed by the use of a sentence will be accounted for more or less in accord with the simple view, so that its propositional content will be determined more or less directly by the meaning of the sentence. However, very often the way in which a speaker is observing the Cooperative Principle and maxims will not be obvious given only the basic locutionary and illocutionary acts that they have performed. Instead, it will be possible to see the speaker as adhering

to the maxims only on the assumption that their basic locutionary and illocutionary acts were shaped by their having certain beliefs and, moreover, by their reasonably intending it to be recognizable, in part by appeal to the assumption that they are observing the maxims, that they have those beliefs.

To see how this might work in a particular case, consider the following gloss on the example given above, involving someone responding to a request for information about a student's philosophical competence by saying, "They have neat handwriting." The audience might reasonably be expected to recognize that they were intended to reason in the following way: the speaker must be being cooperative since, otherwise, they wouldn't have replied to the request. The speaker cannot be unable to say something more relevant, since the question concerns one of their students. The speaker must, therefore, be wishing to impart information that they are reluctant to write down. This supposition is tenable only if the speaker thinks that the student is no good at philosophy. This, then, is what the speaker is implicating. (Gloss modified from Grice, 1989, p. 33. It's worth noting that Grice presented his account of conversational implicature as preliminary and partial and, moreover, as covering only a subclass of non-conventional implicatures. See, e.g., Grice, 1989, p. 26, but compare p. 31. Grice says very little about the class of non-conventional, non-conversational implicatures. Grice's main focus is on implicatures carried by natural language analogues of the logical constants, on which see Edgington, 2006.)

Grice's distinction between basic and derivative illocutionary acts – that is, between what a speaker strictly and literally states and what they implicate by stating it – can help to illuminate three important types of possibility.

The first type of possibility is that the very same basic illocutionary act might serve as a means to the performance of a wide variety of derivative illocutionary acts. For example, in some circumstances, as we've just seen, one might state that a student has neat handwriting in order to communicate skepticism about their philosophical ability. However, in other circumstances – for example, circumstances in which the student is applying for a position as an amanuensis – one might state the same thing in order to communicate their suitability for the role. Thus, the fact that a form of words can be used to perform illocutionary acts with a variety of propositional contents does not alone demonstrate that the form of words is ambiguous, or carries complex meanings that determine different propositional contents in different circumstances.

The second type of possibility is that speakers might be unwilling to state something, and might even judge that stating it would be incorrect, or that what was stated would be false, not because stating it would be incorrect, or because what was stated would be false, but rather because stating it would give rise to implicatures that were false and, so, would be misleading. Suppose, for example, that the applicant for the position of amanuensis in tomorrow's examination had been injured and would not be able to write until the day after tomorrow. In that case, although it would remain true that the applicant has neat handwriting, one might nonetheless be unwilling to state that they have, since doing so would be apt to communicate that one thought them to be suitable for the position. Thus, the fact that in certain circumstances speakers are unwilling assertively to use some words, and even the fact that they would judge the assertive use of those words to be incorrect, or judge what was thereby communicated to be false, does not alone demonstrate that what would be stated in those circumstances would be false.

The third type of possibility is that speakers might be willing to state something, and might even judge that stating it would be correct, or that what was stated would be true, not

because stating it would be correct, or because what was stated would be true, but rather because stating it would give rise to implicatures that were true. For example, one might happily state that the injured applicant is not going to die with the intention of communicating that they will not die as a result of their recent injury. Thus, the fact that in certain circumstances speakers are willing assertively to use some words, and even the fact that they would judge the assertive use of those words to be correct or what was thereby communicated to be true, does not alone demonstrate that what would be stated in those circumstances would be true.

At the most general level, Grice offered an account on which there is no backward road from speakers' uncritical judgments about the correctness or incorrectness of the illocutionary acts performed by the use of some words to facts about the basic illocutionary acts that are so performed. And so, even on the assumption that there is a backward road from the propositional contents of basic illocutionary acts to facts about the meanings of the words used to perform those acts, the route from judgments about illocutionary acts *per se* to meanings is, at best, indirect. It's important to observe, however, that that consequence doesn't arise simply from Grice's account of the distinction between basic and derived illocutionary acts. Rather, it depends in addition on the view that speakers and audiences can sometimes be insensitive to the application of that distinction. Given that the range of illocutionary acts is supposed to be determined by speakers' intentions, and moreover by their intentions to make recognizable their intentions, the required form of insensitivity might reasonably be found puzzling. (See, e.g., Grice, 1989, p. 49; Recanati, 2004, pp. 5–22. Note that this issue appears most pressing when we attend to cases in which what is at issue is the distinction between a basic illocutionary act and derivative illocutionary acts performed on its basis. For in those cases, it is natural to assume that speakers and audiences will need to be cognizant of the intentions with which the various acts are performed. However, if we also allow cases in which a derivative illocutionary act is performed on the basis not of a more basic illocutionary act, but of a locutionary act, we might be more willing to allow that speakers and their audiences can be appropriately insensitive. See Bach, 2001.)

5 The Orthodox View

Grice's distinction between basic and derivative illocutionary acts ruins the simple view, on which the propositional contents of illocutionary acts are determined by the meanings of the words that are used in order to perform those acts. However, it makes available a more nuanced successor, which Grice also sought to defend: the orthodox view. According to the orthodox view, the propositional contents of *basic* illocutionary acts are to a large extent determined by the meanings of the words used to perform them. Variation in the illocutionary acts performed by the use of some words from occasion to occasion is therefore mainly due not to variation in the basic illocutionary acts that are performed, but rather to variation in the derivative illocutionary acts that are performed. More carefully, the orthodox view holds the following: (i) the propositional contents of basic illocutionary acts are to a large extent determined by the contents of locutionary acts; (ii) the contents of locutionary acts are determined by the meanings of the words that are used to perform them in combination with meaning-specified features of the circumstances in which the locutionary acts are performed; and (iii) the contents of locutionary acts determine truth-conditions or satisfaction conditions (where the latter are conditions in which a predicate like "is green"

would be true of an individual). (On the natural assumption that illocutionary act contents determine truth-conditions, (iii) approximates to a corollary of (ii), for locutionary act contents must be truth-conditional if they are to determine the truth-conditional contents of other acts.) According to the orthodox view, there are two sorts of cases in which the contents of basic illocutionary acts are not determined by the meanings of the words used in their performance.

The first sort, which we will henceforth ignore, comprises cases in which speakers advertently or otherwise perform basic illocutionary acts without intentionally exploiting the meanings of the words that they use in order to make recognizable their illocutionary intentions. This might happen if a speaker were intentionally to use a sentence in a way that departed from its meaning, but in such a way that they were nonetheless able to make recognizable their basic illocutionary intentions. Cases of this sort might include slips of the tongue, malapropism, and deliberately creative uses of language. (Davidson, 1986, contains a useful discussion of such cases.)

The second sort comprises cases in which speakers make use of sentences involving *indexicals* – for example, "I," "here," or "now," "She," and "That cat." These are expressions that, despite carrying unitary meanings, vary in the contribution that they make to illocutionary content. For example, "here" is used in order to say something about wherever the speaker happens to be, which might well vary from occasion to occasion. Similarly, "That cat" is used in order to say something about a cat that is distinctively salient, or that is made distinctively salient by an accompanying gesture. In both sorts of case, the meaning of the expression figures in making recognizable the basic illocutionary act that the speaker performs. However, it serves to make recognizable that act only in conjunction with information about the circumstances of speaking – for example, where the speaking takes place, or what the speaker is pointing at. The meanings of such expressions serve not to determine a specific contribution to the contents of illocutionary acts performed by their use, but rather to provide more or less detailed guidance about how information about circumstances is to be used in discerning such a contribution.

We can usefully distinguish two types of indexical, the discretionary and the non-discretionary. Non-discretionary indexicals – for example, "I" or "now" – have meanings that provide very detailed guidance to the ways in which the contents of acts performed by their use are to be discerned. The meanings of such expressions determine a range of relevant circumstantial features and specify the way that those features figure in determining a specific contribution to content. For instance, the meaning of the indexical "I" might be taken to determine a function from a specific feature of the circumstance – that is, the identity of the speaker (or "agent of the context") – to a contribution to truth-conditional content: reference to the speaker. Although the meanings of such expressions do not alone determine a contribution to truth-conditional content, those meanings combine with specific circumstantial features that the meanings select in order to determine such a contribution. Thus, in so far as the speaker intends to perform an act that exploits the meanings of such expressions, no further role is played by their illocutionary intentions either in determining the contents of that act or in selecting specific features of the circumstances that are to be exploited in discerning their illocutionary intentions. Rather, the meanings of the expressions dictate which locutionary acts are performed by their use, and combine with specific features of the circumstances of use in order to determine the contents of those locutionary acts. With respect to the use of non-discretionary indexicals, the orthodox view holds that locutionary content determines illocutionary content.

Discretionary indexicals offer less detailed guidance about how circumstantial features are to be exploited in order to discern illocutionary content. For instance, the meaning of "That cat" provides some guidance to discerning a speaker's illocutionary intentions – indicating, for example, that they intend to say something about a certain salient cat – but leaves it up to speakers how they will exploit, or manipulate, the circumstances of speech in order to make recognizable which cat they intend to say something about. Thus, since meanings alone fail to combine with circumstances in order to determine a truth-conditional content, determination of truth-conditional content for discretionary indexicals can't take place at the level of locutionary content. Rather, since the determination of truth-conditional content depends on speakers' illocutionary intentions, it occurs with respect only to the illocutionary act. With respect to the use of discretionary indexicals, the orthodox view allows that locutionary content fails to determine illocutionary content. Furthermore, the orthodox view allows that in such cases, locutionary content fails to determine truth-conditions, although it will typically still determine satisfaction conditions. For although the locutionary use of "That cat is on the mat" will fail to determine to which cat the speaker intends to refer, it is open to hold that it will determine that the sentence is true of something if and only if it is a cat and is on the mat. (See, e.g., Bach, 1992; Perry, Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS.)

Returning to the general character of the orthodox view, we can now say the following. In cases in which a sentence is used intentionally in accord with the meanings of its constituent words and structure, those meanings figure in two types of way in making recognizable the propositional contents of basic illocutionary acts. First, the meanings of the sentence's indexical constituents serve as guides to the use of circumstantial information in discerning aspects of the speaker's illocutionary intentions. In the case of non-discretionary indexicals, this is a matter of determining the content of the locutionary act performed by their use. In the case of discretionary indexicals, it is a matter of determining the content of an illocutionary act. Second, the meanings of the sentence's remaining constituents, including non-discretionary indexical constituents, serve simply to determine the remaining contribution to locutionary, and so illocutionary, content. According to the orthodox view, then, there are only three ways in which a sentence could be used in accord with its meaning in order, on different occasions, to perform illocutionary acts with different propositional contents. First, the sentence might be used in order to perform derivative illocutionary acts with any of a variety of contents, as long as that variety were explicable by appeal to a combination of the performance of a basic illocutionary act together with the Cooperative Principle (or principles of broadly the same sort). Second, a sentence might be used in order to perform basic illocutionary acts with a variety of contents if the sentence were to involve indexicals. Third, a sentence might be used in order to perform illocutionary acts with a variety of contents if the sentence were ambiguous, so that it could be used in accord with different of its meanings on different occasions.

As noted in the introduction, attempts to sharpen the division between semantics and pragmatics have been guided by a number of considerations. The first and most important consideration is the idea that semantics should concern information about some of the stable, broadly conventional properties of words and sentences: the meanings of those words and sentences. Such information is of a sort that can be learned in advance and then brought to bear in seeking to understand particular uses of those words and sentences. The second consideration is that semantics should figure in explaining what speakers are able to state, or ask, or command, by the use of words and sentences – that is, it should have something to say about

the propositional contents of basic illocutionary acts. The third consideration, which may be viewed as a consequence of the assumption that the propositional contents of speech acts determine truth-conditions, is that semantics should have something to say about truth-conditions. The orthodox view allows for a division of labor between semantics and pragmatics that comes close to respecting all three considerations. For according to the orthodox view, the stable properties of words and sentences for the most part determine the propositional contents of locutionary acts that are performed by their use. And the contents of those locutionary acts for the most part determine the contents of basic illocutionary acts. In so far as the determination of illocutionary content is due only for the most part to the contents of locutionary acts, that is due to the operation of discretionary indexicals, and so leaves open that in cases in which locutionary act contents fail to determine truth-conditions, they nonetheless determine satisfaction conditions. For although the meaning of 'That is green' leaves it up to speakers' discretion what 'that' is used to refer to, it might be held that 'is green' is true of any such thing (or is satisfied by that thing) if and only if it is green. So, although the orthodox view treats the effects of discretionary indexicals as falling partly within the remit of pragmatics, rather than semantics, it accords to meaning, and so semantics, responsibility for all other effects on locutionary and illocutionary act contents, including the potential effects of non-discretionary indexicals.

Although the orthodox view is widely endorsed, in recent years it has come under increasing pressure from the defenders of views on which the connection between meaning and the propositional contents of basic illocutionary acts is taken to be less straightforward. As we noted, discretionary indexicals differ from non-discretionary indexicals in that they do not fully specify a function from specific features of circumstances of use to a contribution to propositional content, but instead provide only more or less detailed guidance to the discernment of speakers' illocutionary intentions. Opponents of the orthodox view seek to treat that sort of model as more generally applicable: in general, the meanings of words don't determine a contribution to the contents of basic illocutionary acts, but rather provide guidance to that contribution. (Importantly, their claim is that substantive expressions are in that specific way like discretionary indexicals, not that they simply are discretionary indexicals. For one important difference, 'is green' seems to provide more detailed guidance as to the kinds of things that can be stated by its use than do expressions like 'that'.) By contrast with the orthodox view, the opposing view holds that substantive word meaning rarely, if ever, provides sufficient guidance to facilitate recognition of speakers' basic illocutionary intentions. Speakers must therefore exploit features of the circumstance of speaking over and above the meanings of the words and sentences that they use in order to make recognizable their basic illocutionary intentions. On this view, the tasks of making one's basic illocutionary intentions recognizable, and of recognizing another's basic illocutionary intentions, will often exploit the sorts of information and principles that figure on the orthodox view only in mediating the transition from basic to derived illocutionary acts. Furthermore, just as discretionary indexicals fail to determine a contribution to truth-conditional content, so word meanings in general fail to determine such a contribution. Truth- and satisfaction conditions are determined not at the level of meaning or locutionary act, but only at the level of illocutionary act.

One main motivation for opposition to the orthodox view comes from the fact that our judgments about what sentences can be used to state in particular circumstances seem to vary in ways that don't obviously fit the orthodox view. That is, our pattern of judgments isn't easily accounted for by appeal either to ambiguity, or to the operation of ordinary indexicals or demonstratives, or to distinctions between the contents of basic and derived illocutionary acts. In the following two sections, we'll briefly consider an argument of this

sort against the orthodox view due to Charles Travis (Chapter 6, PRAGMATICS). Connectedly, we observed earlier that interesting applications of the distinction between basic and derivative illocutionary acts rely upon treating competent speakers as somewhat insensitive to the way the distinction applies. Opposition to the orthodox view has been bolstered by an unwillingness to treat the pattern of competent speakers' judgments as due to an insensitivity to the sorts of distinctions between basic and derivative illocutionary acts that would be required in order to preserve the orthodox view. (Related attacks on the orthodox view may be found in Austin, 1962a; Bach, 1994; 2001; Carston, 2002; Neale, 2005; Pietroski, 2003; 2005; Recanati, 2004; 2010; Sperber and Wilson, 1995; Travis, 2008.)

6 Occasion-Sensitivity

In the previous section, we saw that the orthodox view is committed to the following claims:

(OV1) Most substantive expressions, other than explicit discretionary indexicals, do not function in a similar way to discretionary indexicals.

(OV2) With respect to those substantive expressions, meaning combines with features of circumstances of speech to determine the content of speakers' locutionary acts in a way that doesn't depend on speakers' further illocutionary intentions.

(OV3) With respect to the use of such expressions, locutionary act content determines basic illocutionary act content.

(OV4) Basic illocutionary act content determines truth-conditions.

(OV5) With respect to the use of such expressions, locutionary act content determines truth-conditions.

(OV4), the claim that basic illocutionary act content determines truth-conditions, is common ground between Travis and proponents of the orthodox view. Travis aims to provide reasons to reject (OV3) and (OV5). He does so by attempting to show that the truth-conditions determined by basic illocutionary acts vary in a way that cannot be explained by appeal to the possession of truth-conditions by locutionary acts or by the role of explicit indexicals. If successful, Travis's argument would in the first place undermine (OV3), by showing that the connection between locutionary act content and basic illocutionary act content is in general – not only in cases involving explicit discretionary indexicals – less straightforward than the proponent of the orthodox view maintains. Contrary to (OV1), most substantive expressions would function in a similar way to discretionary indexicals: their meanings would provide more or less detailed guidance to discerning speakers' illocutionary intentions. In the second place, Travis's rejection of (OV3) would put pressure on (OV5). One source of pressure here is the apparent dependence of (OV5) on (OV4): if the only reason for holding that locutionary act contents determine truth-conditions were that those contents determine illocutionary act contents that determine truth-conditions, then that reason would lapse with rejection of (OV3). A further, connected source of pressure is that assessment as to truth seems typically to be directed at the contents of illocutionary acts. We typically seek to ascertain, for example, whether what someone states is true or false. If there were cases in which locutionary act contents determined illocutionary act contents, as per (OV3), then

truth assessments of illocutionary acts could in those cases be used in order to gain insight into the truth-conditions determined by their corresponding locutionary acts. However, in the absence of that connection between locutionary and illocutionary acts, it would be difficult to see how to gain purchase on the truth-conditions supposedly determined by locutionary act contents. Minimally, Travis's argument presents the defender of a truth-conditional conception of meaning or locutionary content with the challenge of providing grounds other than the combination of (OV3) and (OV4) for endorsing their conception.

Travis develops an argument against the orthodox view based around cases like the following:

> Pia's Japanese maple is full of russet leaves. Believing that green is the color of leaves, she paints them. [Case 1:] Returning, she reports, 'That's better. The leaves are green now.' She speaks truth. [Case 2:] A botanist friend then phones, seeking green leaves for the study of green-leaf chemistry. 'The leaves (on my tree) are green,' Pia says. 'You can have those.' But now Pia speaks falsehood. (Travis, Chapter 6, PRAGMATICS)

According to Travis, we can suppose that the same words, with the same meanings, are used in both cases. Moreover, we can assume that the operations of discretionary and explicit non-discretionary indexicals are confined to determining which leaves are being spoken about, and that the same leaves are being spoken about in both cases. Thus, we can assume that the same locutionary act, with the same locutionary content, is performed in case 1 and case 2. And since the same leaves, in the same condition, are being spoken about in both cases, if 'The leaves are green' were used in both cases to make the same statement and, so, in accord with (OV4) to determine the same truth-conditions, then Pia would either state a truth in both cases, or state a falsehood in both cases. Since it is plausible that Pia can state a truth (and no falsehood) in case 1 and that she can state a falsehood (and no truth) in case 2, even though the leaves are in the same condition in both cases, we have that 'The leaves are green' is not used in both cases to make the same statement. So, we have that the statement that Pia makes in case 1 is not the same as the statement that she makes in case 2. Finally, if we assume that the statements that Pia makes in both cases are the contents of basic illocutionary acts, we have grounds for rejecting (OV3). For we have that Pia performed locutionary acts with the same contents and basic illocutionary acts with different contents and, moreover, that the difference in illocutionary content was not due to the operation of explicit discretionary indexicals. And since *n*-tuples of cases of this sort can be constructed for most, if not all, substantive expressions, there are grounds for generalizing Travis's conclusion: most, if not all, substantive expressions function in a similar way to discretionary indexicals.

Travis's argument depends upon three main assumptions. The first assumption is that the variation in truth-conditions exhibited across cases 1 and 2 affects basic illocutionary acts performed in those cases. For if the variance affected only derivative illocutionary acts, it would be consistent with maintaining (OV3). The second assumption is that the variation is not due to ambiguity. For if 'The leaves are green' was used with relevantly different meanings across the two cases, the variation in truth-conditions might be traced to a difference in locutionary act content. The third assumption is that the variation is not due to the operation of non-discretionary indexicality. For if it were due to such indexicality, then the variation in truth-conditions might again be due to variation in locutionary act content. Travis's second and third assumptions have been subjected to interesting challenges in

recent years. However, no clear case has yet been presented in favor of the view that the variation in truth-conditions that Travis highlights can be accounted for by a combination of non-discretionary indexicality and ambiguity. (For approaches that appeal to indexicals, see Rothschild and Segal, 2009; Stanley, 2000; 2002; 2005; Stanley and Szabó, 2000; Szabó, 2001. For an account that appeals in addition to ambiguity, see Kennedy and McNally, 2010. For discussion see Clapp, 2012; Collins, 2007. The main weakness exhibited by extant proposals is that no attempt has been made to provide evidence for a plausible account of a function from meaning-specified features of circumstances of speech to truth-conditional contents. Thus, such accounts seem, at best, either to mitigate discretionary variation without eradicating it, or to resolve into versions of Travis's position. Versions of that complaint are developed in Clapp, 2012, and Rothschild and Segal, 2009.)

7 Basic and Derivative Illocutionary Acts

As we saw in discussing Grice's distinction between basic and derivative illocutionary acts, theorists have been willing to allow that competent speakers can be insensitive to the distinction between what speakers state and the contents of derivative illocutionary acts that speakers thereby perform. If defensible, such an allowance would make space for the following response to Travis's argument. Although there is variation across cases of the sort that Travis emphasizes, that variation affects only the contents of derivative illocutionary acts. In so far as we are inclined to judge that there is variation in the contents of what speakers state in the cases that Travis presents, that is a mistake fostered by our insensitivity to the distinction between what speakers state and the derivative illocutionary acts that they perform on the basis of so stating.

The central difficulty with this position arises because illocutionary acts of stating embody intentionally undertaken commitments. Let's suppose that the locutionary act performed by use of 'The leaves are green' determines the content of a basic illocutionary act. That act is therefore performed in both case 1 and case 2. Since the content of that act determines truth-conditions, it will represent a harmless simplification to suppose that those truth-conditions are such as to make the illocutionary act false in both cases. To a first approximation, the illocutionary act has a content that would be true if and only if the leaves were naturally green. Thus, let's suppose that in case 2 Pia uses 'The leaves are green' in order to perform a basic illocutionary act with that content. As Travis maintains, and as Pia could easily be brought to accept, she thereby states a falsehood. However, contrary to Travis's description, Pia performs the same basic illocutionary act in case 1, thereby committing herself to something false in that case too, her protestations to the contrary notwithstanding. According to the position currently being considered, although Pia commits to a falsehood in case 1, she thereby also performs a derivative illocutionary act with a true content and thereby commits to something true. Furthermore, Pia's failure to recognize, or to accept, that she commits to a falsehood in case 1 is to be explained by her insensitivity to the distinction between basic and derivative illocutionary acts. That is, it is explained by her failure to recognize that in addition to committing herself to a truth by performance of the derivative illocutionary act, she committed herself to falsehood by performing the basic illocutionary act. The difficulty, then, is that it is hard to accept that in case 1 Pia openly and intentionally commits herself to a falsehood, even though that is something that she failed to recognize and would even be prepared to deny. That consequence is hard to accept

because it conflicts with the natural presumption that agents know the intentions with which they act. Furthermore, not only would Pia fail to recognize that she had openly and intentionally committed herself to a falsehood in case 1, but her audience would also fail to recognize it. And that consequence is in tension with the requirement that Pia should act with the reasonable expectation that her illocutionary intentions will be recognizable. (See, e.g., Bach, 1994; 2001; Soames, 2005.)

It's important to emphasize that the claim that Pia does not perform the same basic illocutionary act in cases 1 and 2, and so does not make the same statement, is compatible with allowing that she performs other types of act in both cases. As we saw, Travis allows that Pia performs the same locutionary act in the two cases. Thus, on a broadly locutionary understanding of what speakers say, as opposed to the illocutionary understanding that we've characterized in terms of what speakers state, Travis can allow that Pia says the same thing in both cases. (See, e.g., Bach, 2001; Travis explicitly allows that the same thing might be said in pairs of cases like those involving Pia in his 2008, p. 156.) All that Travis is committed to denying, as a consequence of his denial of (OV5), is that what speakers say – on the operative locutionary understanding of what they say – determines truth or satisfaction conditions. For that reason, arguments designed to show that Pia says the same thing in cases 1 and 2 are impotent unless supplemented with reason to think that what she says in both cases determines a single set of truth-conditions. (Compare Cappelen and Lepore, 2005; Cappelen and Hawthorne, 2009.)

In summary, then, the defender of the orthodox view has two options. First, they might try to show that Pia performs the same basic illocutionary act in cases 1 and 2. That would be a way of seeking to defend (OV3) and (OV5). In doing so, they would need to provide evidence not only that Pia says the same thing in the two cases, but that she makes the same statement, and thereby takes on the same commitments. They would thereby incur the burden of explaining Pia's insensitivity to the commitments that she thereby takes on. Second, they might try to argue that there are locutionary acts that Pia performs in both cases and that those acts have contents that determine truth-conditions. That would be a way of conceding (OV3) while seeking to defend (OV5). Since such acts are not illocutionary acts, and so are not determined by Pia's reflexive intentions, it is liable to be easier to explain why we are insensitive to their occurrence. However, for the same reason, it will be harder to defend the claim that such acts have contents that determine truth-conditions, since it will no longer be possible to use the connection between illocutionary acts and truth-conditions (as per (OV4)) in order to do so.

8 Conclusion

We began with the idea that semantics concerns the stable meanings of words and expressions while pragmatics concerns language use, or things done with words. We saw that that rough distinction is associated with other concerns, including a concern to understand the features of language and its use that are responsible for what speakers state, and for the determination of truth-conditions. According to the orthodox view, the initial distinction between stable meanings and use is connected with what is stated and with truth-conditions. For according to the orthodox view, word meanings combine with circumstances to determine the contents of locutionary acts, and locutionary acts combine with illocutionary intentions with respect to explicit discretionary indexicals in order to determine the contents of basic

illocutionary acts, including acts of stating. And illocutionary acts, including acts of stating, have contents that determine truth-conditions. Thus, there is a more or less direct connection between meanings, the things that speakers state, and truth-conditions.

We saw that the orthodox view comes under pressure from reflection on certain forms of variation in the illocutionary acts that speakers perform. According to the opposing view, personified here by Travis, the connection between meaning and basic illocutionary act contents – in particular, the things speakers state – is less direct, and is mediated by some of the same sorts of factors that the orthodox view exploits in order to connect basic and derived illocutionary acts. We considered, and provided some grounds for rejecting, a defense of orthodoxy that sought to treat the variations that Travis highlights as occurring only with respect to derivative illocutionary acts. However, deciding the outcome of the dispute over the standing of the orthodox view will require further work. In particular, it will require further work on the nature of illocutionary acts, the nature of illocutionary intentions, and the extent to which speakers and their audiences can be expected to be cognizant of the illocutionary acts that speakers perform.

Note

1 I'm grateful for comments to Ariane Beeston, Giulia Felappi, Bob Hale, Simon Hewitt, Hemdat Lerman, Eliot Michaelson, Alexander Miller, Daniel Morgan, and Mark Textor.

References

Austin, J. L. 1962a. *Sense and Sensibilia*. Reconstructed from the manuscript notes by G. J. Warnock. Oxford: Oxford University Press.

Austin, J. L. 1962b. *How to Do Things with Words*. Oxford: Clarendon Press. 2nd edn, edited by M. Sbisà and J. O. Urmson. Oxford: Oxford University Press, 1975.

Bach, K. 1992. "Intentions and demonstrations." *Analysis*, 52: 140–146.

Bach, K. 1994. "Conversational impliciture." *Mind & Language*, 9(2): 124–162.

Bach, K. 1999. "The semantics–pragmatics distinction: what it is and why it matters." In *The Semantics/Pragmatics Interface from Different Points of View*, edited by K. Turner, pp. 65–84. Oxford: Elsevier.

Bach, K. 2001. "You don't say?" *Synthese*, 128(1): 15–44.

Bach, K., and R. Harnish. 1979. *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Bird, G. 1981. "Austin's theory of illocutionary force." In *Midwest Studies in Philosophy*, vol. 6, edited by P. A. French, T. E. Uehling Jr, and H. K. Wettstein. Minneapolis: University of Minnesota Press.

Cappelen, H., and J. Hawthorne. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cappelen, H., and E. Lepore. 2005. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

Carston, R. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Clapp, L. 2012. "Three challenges for indexicalism." *Mind & Language*, 27(4): 435–465.

Collins, J. 2007. "Syntax, more or less." *Mind*, 116(464): 805–850.

Davidson, D. 1986. "A nice derangement of epitaphs." In *Philosophical Grounds of Rationality*, edited by R. Grandy and R. Warner, pp. 157–174. Oxford: Clarendon Press.

Edgington, D. 2006. "The pragmatics of the logical constants." In *The Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. C. Smith, pp. 768–793. Oxford: Oxford University Press.

Geach, P. T. 1965. "Assertion." *Philosophical Review*, 74(4): 449–465.
Grice, P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Hornsby, J. 1988. "Things done with words." In *Human Agency: Language, Duty, and Value*, edited by J. Dancy, J. M. E. Moravcsik, and C. C. W. Taylor, pp. 27–46. Stanford, CA: Stanford University Press.
Hornsby, J. 1994. "Illocution and its significance." In *Foundations of Speech Act Theory*, edited by S. L. Tsohatzdis, pp. 187–207. London: Routledge.
Hornsby, J. 2006. "Speech acts and performatives." In *The Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. C. Smith, pp. 893–909. Oxford: Oxford University Press.
Kennedy, C., and L. McNally. 2010. "Color, context, and compositionality." *Synthese*, 174(1): 79–98.
McDowell, J. 1980. "Meaning, communication, and knowledge." In *Philosophical Subjects*, edited by Z. van Straaten, pp. 117–139. Oxford: Oxford University Press.
Morris, C. 1938. *Foundations of a Theory of Signs*. Chicago: University of Chicago Press.
Neale, S. 2005. "Pragmatism and binding." In *Semantics versus Pragmatics*, edited by Z. G. Szabó, pp. 165–285. Oxford: Clarendon Press.
Pietroski, P. 2003. "The character of natural language semantics." In *Epistemology of Language*, edited by A. Barber, pp. 217–256. Oxford: Oxford University Press.
Pietroski, P. 2005. "Meaning before truth." In *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, edited by G. Preyer and G. Peter, pp. 255–302. Oxford: Clarendon Press.
Recanati, F. 2004. *Literal Meaning*. Cambridge: Cambridge University Press.
Recanati, F. 2010. *Truth-Conditional Pragmatics*. Oxford: Clarendon Press.
Rothschild, D., and G. Segal. 2009. "Indexical predicates." *Mind & Language*, 24(4): 467–493.
Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge University Press.
Soames, S. 2003. *Philosophical Analysis in the Twentieth* Century, vol. 2, *The Age of Meaning*. Princeton, NJ: Princeton University Press.
Soames, S. 2005. "Naming and asserting." In *Semantics versus Pragmatics*, edited by Z. G. Szabó, pp. 356–382. Oxford: Clarendon Press.
Sperber, D., and D. Wilson. 1995. *Relevance: Communication and Cognition*, 2nd edn. Oxford: Blackwell.
Stanley, J. 2000. "Context and logical form." *Linguistics and Philosophy*, 23(4): 391–434.
Stanley, J. 2002. "Making it articulated." *Mind & Language*, 17(1–2): 149–168.
Stanley, J. 2005. "Semantics in context." In *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, edited by G. Preyer and G. Peter, pp. 221–253. Oxford: Clarendon Press.
Stanley, J., and Z. G. Szabó. 2000. "On quantifier domain restriction." *Mind & Language*, 15(2–3): 219–261.
Strawson, P. F. 1949. "Truth." *Analysis*, 9(6): 83–97.
Szabó, Z. G. 2001. "Adjectives in context." In *Perspectives on Semantics, Pragmatics, and Discourse*, edited by R. Harnish and I. Kenesei, pp. 119–146. Amsterdam: John Benjamins.
Szabó, Z. G. 2006. "The distinction between semantics and pragmatics." In *The Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. C. Smith, pp. 361–389. Oxford: Oxford University Press.
Travis, C. 2008. *Occasion-Sensitivity: Selected Essays*. Oxford: Oxford University Press.

6

Pragmatics

CHARLES TRAVIS

Here are two non-equivalent characterizations of pragmatics. Pragmatics (first version) concerns the linguistic phenomena left untreated by phonology, syntax, and semantics. Pragmatics (second version) is the study of properties of words which depend on their having been spoken, or reacted to, in a certain way, or in certain conditions, or in the way, or conditions, they were (Kalish, 1967).

Here are two equally non-equivalent characterizations of semantics. Semantics (first version) is, by definition, concerned with certain relations between words and the world, and centrally with those on which the truth or falsity of words depends: thus David Lewis's slogan, "Semantics with no treatment of truth conditions is not semantics" (Lewis, 1972, p. 169). Semantics (second version) is defined by this idea: "A theory of meaning for a language should be able to tell us the meanings of the words and sentences which comprise that language" (Platts, 1980, p. 2). So what a semantic theory of English, say, must do is, for each English expression, provide a specification of what it means. Semantics in general would be an account of the nature of such particular theories, or of their subject-matter.

Combine these different ideas, and you get a substantial thesis: such things as English sentences *have* statable conditions for truth, and meanings can be given in or by stating these. That *might* be wrong. Perhaps, as J. L. Austin suggested, questions of truth arise at a different level entirely from that of expressions of a language. Perhaps conditions for truth depend, pervasively, on the circumstances in which, or the way in which, words were produced. If so, then on the second version of pragmatics and the first version of semantics, semantic questions are pragmatic ones; whereas semantics (second version), however it is to be done, would have little or nothing to do with truth-conditions. Call this the pragmatic view.

This chapter argues that the pragmatic view is the right one; that it is intrinsically part of what expressions of (say) English mean that any English (or whatever) sentence may, on one speaking of it or another, have any of indefinitely many different truth-conditions, and that any English (or whatever) expression may, meaning what it does, make any of many

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

different contributions to truth-conditions of wholes in which it figures as a part. I will first set out the reasons for thinking so, then discuss a few of the most significant consequences.

The issue also emerges in asking what words are for. On one view, bracketing ambiguity, indexicals, and demonstratives (see Chapter 38, THE SEMANTICS OF PRAGMATICS AND INDEXICALS), for each declarative English sentence there is a thought which is the one it expresses; its role in English is to express that one. On the pragmatic view this is just what is not so. Independent of ambiguity, indexicality, and so on, what meaning does is to make a sentence a means for expressing *thoughts* – not some *one* thought, but any of myriad different ones. Meaning does that in making a sentence a particular description of how things are, so a means for describing things as that way. Any description admits of many different applications. The same description, applied differently, yields different thoughts. A right application, where there is one, is fixed by circumstances of producing the description, not just by the description itself. If a sentence may thus equally well express any of many thoughts, conditions for the truth of one of these cannot be conditions for the truth of the *sentence*.

1 Semantic Properties

There are properties words have, and would have, no matter how we understood them. Being spoken loudly or at 3 p.m. are two. Then there are properties words have, or would have, on one understanding of them, but would lack on another – properties words have, if at all, only in virtue of their being rightly understood in the way they are. I want to consider two classes of such properties.

The first sort of property is one of relating in a given way to truth (or falsity). Properties of being true (false) if, given, of, or only if, thus and so, or thus, or the way things are, are all within this class. (They are all properties words might have on one understanding, and lack on another.) For future convenience, I exclude being true or false *simpliciter* from this class, though I include being true (false) given the way things are. I call these properties *truth-involving*, and any set of them a *truth-condition*.

The second sort are properties identified without mention of truth, and on which truth-involving properties depend. Such properties include such things as describing X as Y, calling X Y, saying X to be Y, and speaking of X. The words 'is red,' for example, speak of being red and, on a speaking, may have called something red. These properties identify *what* words say. I will call them *content-fixing,* and any set of them a *content*.

One might wonder whether content-fixing properties are not really truth-involving ones in disguise – whether, for example, to call something red is not just to say (of it) what is true of such-and-such things, and true of a thing under such-and-such conditions. In what follows, we will find out whether that is so.

The properties indicated so far might reasonably be called semantic, not worrying overly for the moment about boundaries between syntax and semantics. I will call them that, and any set of them a *semantics*. The latitude allowed here means that not every semantics in the present sense is one words might have. Some semantic properties may exclude others. Calling something a fish, for example, may exclude, *tout court*, saying what is true of my piano. Call a semantics some words might have *coherent*, keeping in mind that a semantics might thus be coherent on some occasions for speaking, while not on others.

We can raise questions about a semantics, or sort of semantics, without saying which items might have it – whether, for example, English sentences or something else might do so. One thing we may ask of a given semantics is whether it *requires* any further semantics – whether there is a semantics which any words with it must have. Or we may ask whether it is supplementable in a variety of – perhaps mutually exclusive – ways; whether words with it may, for all that, have any of various further semantics.

It is interesting to ask, in particular, whether the semantics an English sentence has in meaning what it does is compatible with any of many supplementations, specifically with any of a variety of truth-conditions. To answer that we need not first say what semantics meaning does confer. We need only find a number of speakings of the sentence on each of which it had whatever semantics its meaning does confer; on each of which, as much as any of the others, those words *did* mean what they do mean. In specific cases we may convince ourselves of that much without knowing just *which* properties meaning confers.

2 The Pragmatic View

Is what a sentence means compatible with semantic variety – specifically variety in truth-conditions – across its speakings? Consider this sentence:

(1) The leaves are green.

The words 'are green,' meaning what they do, are means for calling things green. Similarly, meaning what they do, 'The leaves,' when spoken as in (1), purport to speak of some leaves. What its (present) tense means makes (1), on a speaking, purport (roughly) to speak of things at the time of that speaking. Consider speakings of (1) in which the words did all this, and in all other respects (if any) meant what they *mean*. Does that much semantics require them to have just *one* full semantics on all such speakings? Or is that much compatible with semantic variety, and, specifically, with those words having, on different speakings, any of many truth-conditions?

A story. Pia's Japanese maple is full of russet leaves. Believing that green is the color of leaves, she paints them. Returning, she reports, 'That's better. The leaves are green now.' She speaks truth. A botanist friend then phones, seeking green leaves for a study of green-leaf chemistry. 'The leaves (on my tree) are green,' Pia says. 'You can have those.' But now Pia speaks falsehood.

If the story is right, then there are two distinguishable things to be said in speaking (1) with the stipulated semantics. One is true; one false; so each would be true under different conditions. That semantics is, then, compatible with semantic variety, and with variety in truth-involving properties. So what the words of (1) mean is compatible with various distinct conditions for its truth.

But is the story right? There are just two grounds for rejecting it. First, one might reject its data by claiming that both speakings of (1), above, share a truth-value, require the same for truth, and are true of the same. Second, one might accept the phenomena as presented, but claim that they *are* accounted for by what (1) means – either by some ambiguity in (1), or by some particular way in which what (1) means makes what it says depend systematically on the circumstances of its speaking.

Consider the first option. Either the stipulated semantics makes (1) true of painted leaves, or it makes (1) false of them, *punkt*. If one of these disjuncts is right, appearances to

the contrary may be explained in any of a variety of ways. The first task, though, is to choose. Which disjunct is right? One must choose in a principled way. What the words mean must make one or the other disjunct plainly, or at least demonstrably, true.

What we know about what words mean will not solve this problem of choice. Nothing we know about what '(is) green' means speaks to this question: If an object is painted green, should its color count as what it would be without the paint, or rather as what it has been colored by painting it? Nor is it plausible that some further development in natural science might resolve this issue. So, it seems, the first option must be rejected. Nor, as we shall see, are colors an unfair example. There are similar problems for any simple predicate, ones left unsolved by what the words in question mean.

We must, then, begin on the second option. Its simplest version is that (1) is ambiguous, or that the words 'are green' are: in one of their senses, they are true of leaves painted green, in another, false of leaves merely painted green. Does 'is green' have such senses in English? I do not think so. But there is a more important question. Suppose it does. Would that yield a different answer to our question about semantic variation?

It would change the answer if the only occasion for saying both true and false things of given leaves in speaking (1) were in case they were painted. But there are indefinitely many more occasions for saying either of two distinct things than that provides for. Suppose the leaves were not painted (or were painted red), but had a fluorescent green mould growing on them. Or suppose they are painted, but in pointillist style: from a decent distance they look green, but up close they look mottled. Is that a way of painting leaves green? It might sometimes, but only sometimes, so count. So there would be two distinct things to be said in the presumed 'paint counts' sense of 'is green.' And so on.

The above need not be the *only* ambiguity in the English 'is green.' But if words are ambiguous in English, there must be a way of saying just what these ambiguities are; so a fact as to how many ways ambiguous they are. The pair of speakings we considered differed in that each invoked a different understanding of what it would be for leaves to be green. There is no reason to think that there is any limit to possible understandings of *that*, each of which might be invoked by some words which spoke on that topic. There is not only an understanding on which painting might make it so, but also one on which painting might make it, so as long as this is not in too loose a pointillist style, or too shiny. And so on, *ad infinitum*. If 'green' has, say, 13 senses, there are, for each of them, various possible (and invokable) understandings of what it would be for leaves to be green in *that* sense. If so, then ambiguity is not a way of avoiding the present conclusion.

It is sometimes said: there is no *uniform* standard for things being green; it is one thing for an apple to be green, another for a tomato to be green, and so on. That idea, though, gets nowhere with the present problem. Throughout, the question has been what it is true to say of *leaves*.

Finally, it might be said that the phenomena show 'green' to be a vague term. Perhaps it is in some sense, though we have so far seen no more reason to say so than there is to say the same of any term. But it is hard to see how vagueness is to the point. In one sense, perhaps, words are vague if there is not enough in a correct understanding of them for deciding whether, given the way the things they speak of are, they ought to count as true or false. The *English* sentence (1) is certainly in that condition. But one *speaking* of it may clearly state what is true, while another clearly states what is false. That can only be so if the semantics of (1) on some speakings of it is substantially richer than that fixed for it by the meanings of its constituents, and richer in different ways for different such

speakings. So what (1) says on a speaking, of given leaves, and so on, is not determined merely by what it, or its parts, mean.

I take the English sentence (1) to illustrate, in the respects noted, what is generally so of a language's sentences – indeed, to illustrate how a sentence of a language *must* function. I have no space for more examples; nor for a satisfying account of why that should be.[1] The reader might anyway test the claim with some further examples of his or her own.

3 Domestications

The above, if correct, answers the initial question: what a sentence means, or what its parts do, is compatible with semantic variety; with variety in what such words say or said, and with variety in their truth-involving properties. One might think that compatible with the traditional view, in which semantics is both the study of what words mean and, centrally, of the conditions for their truth; that all said so far is consistent with the meanings of words determining the conditions for their truth; and even that the general point has long been recognized. One might still think, in other words, that the point may be domesticated within a framework in which what words mean still fixes, in an important sense, what they say wherever spoken. I will discuss two plans for such domestication.

The first plan turns on the idea of ellipsis: some words are to be understood as short for others. A particular 'He'll come,' for example, may be rightly construed as a shortened 'He'll come to the party.' Assuming ellipsis were pervasive, how might it help? If (1) may be used to say any of many things, it must, on different speakings, be elliptical for different things: on each it says what that for which it is then elliptical would say. For this explanation to domesticate the phenomena, the things for which (1) is elliptical must not themselves exhibit semantic variation of the sort that (1) did. For example, if a given instance of (1) is elliptical for 'The leaves are green beneath the paint,' there must not be more than one thing to be said in *those* words. If the phenomena are as I suggest, this assumption is wrong. I leave this suggestion at that.

The second suggestion revolves around this idea: what words mean *does* determine what they say.[2] But it does not do so *simpliciter*. Rather, it does so as a function of some set of factors, or parameters, in speakings of the words. The parameters allow for different things to be said in different such speakings. Such was always in the plan for linking sentences with truth-conditions.

The plan is illustrated by Frege's treatment of the present tense. Frege notes that a speaking of (1) in July might be true, while one in October was false.[3] He observes, correctly, that different things would have been said in each such speaking. One thing this shows is that the tensed verb refers to a specific time or interval, and different ones on different speakings; the words say the leaves to be green at that time.

Frege thought that more was shown. First, that for the present tense the time referred to is always the time of speaking. Second, that where present-tense words are spoken, there is a factor – the time they were spoken – and a function, fixed by what they mean, from values of it to the time they spoke of; in fact, the identity function. So third, that what (1) means determines a function from variables in its speakings to thoughts expressed on those speakings.

Frege's view might be generalized. What *some* words say, or contribute to what is said in using them, varies across speakings of them. Where this is so, the meaning of the words

does two things. First, it determines on just what facts about a speaking the semantic contribution of the words so spoken depends. Second, it determines just how their semantics on a speaking depends on these facts. Specifically, it determines a specifiable function from values of those factors to the semantics the words would have, if spoken where those values obtain.

The above is a hypothesis. *If* it is true, then while the words (1) may say different things on different speakings, what those words mean determines how they so vary. It determines that the words say thus and so where such-and-such factors take on such-and-such values, for any values those factors may take on (where the thus and so said is what would be true under such-and-such conditions). If that is so, it is reasonable to say that what words mean determines what they say, and when they, or that, would be true. It does so by determining effectively how other facts about their speaking matter to such questions.

But is the hypothesis true? First note that semantics is not history. Sentence (1) will have been spoken only a finite number of times before the heat death of the universe. Suppose that each such time something in particular was said. Then, of course, there is a function from parameters of *those* occasions to what was said in (1) on them. There are many such functions, from many such parameters. That is not semantics. What we wanted to know was: if you spoke (1) on such-and-such occasion (as may or may not actually be done), what *would* you say? The question was whether what (1) means provides an answer to that. The historical remark about actual occasions does nothing towards showing that it does.

The point was that the words 'is green,' while speaking of being green, may make any of many semantic contributions to wholes of which they are a part, different contributions yielding different results as to what would count as things being as they are said to be. Are there parameters in speakings of those words which determine just which semantic contribution they would make when? Is there a function such that for each assignment of values to those parameters, there is one particular contribution the words *would* inevitably make, spoken where those values hold? I will not demonstrate here that there are no such things. But there need not be: perhaps for any set of parameters, further possible factors would yield more than one distinguishable thing to be said for fixed values of those.

There are several respects in which the present phenomena are *unlike* central cases where the parameter approach seems promising. One difference is this. In central cases, such as 'I' and 'now,' pointing to given parameters seems to be a part of the terms' meaning what they do. It is part of the meaning of 'I,' and its use in English, that it is a device for a speaker to speak of himself. That suggests speakers as a relevant parameter. If there is no unique semantic contribution, 'I' makes for a fixed value of that parameter, the meaning of 'I' fixes no function from *that* to contributions made in speaking it. By contrast, it is not part of what 'green' means, so far as we can tell, that speakings of it speak of, or refer to, such-and-such parameters. If its contribution, on a speaking, to what is said is a function of some parameters – say, implausibly,[4] the speaker's intentions – saying so is not part of saying what 'green' means. The parameter approach does not *automatically* suggest itself here as it did with 'I.'

This difference between 'I' and 'green' shows up when it comes to saying what was said. Consider a speaking of the words 'I am in Paris.' Ignore any possibilities for various contributions by 'in Paris,' or by the present tense at a time. Then, knowing nothing more about the speaking, we know that, in it, it was said *that* the speaker, whoever s/he may be, was, at the time of speaking, whenever that was, in Paris. However in the dark we may be on those points, we *do* thus specify which fact (or non-fact) was stated. Not so for speakings of (1).

Suppose that Pia spoke those words, and that *we* say of that, 'Pia said that the leaves she spoke of were, at the time of speaking, green.' We will not have said *what* Pia stated unless *our* 'green' made some definite contribution to what *we* said about Pia. But, as we have seen, 'green' may make any of many contributions of the needed sort. If it made one such in *our* words and a different one in Pia's then what we said about her is *false*. We may, for example, have said her to say what would be false of green-painted leaves, while what she said would be true of that. The information contained in the meanings of the words she used is thus not enough for specifying, however uninformatively, *which* fact (or non-fact) she stated.

In speaking (1) literally, one does what then counts as calling leaves green. That may be one thing that sometimes counts as 'saying that the relevant leaves were green.' But *such* a use of 'say that,' if there is one, does not purport to specify which fact (or non-fact) was stated. It says nothing that allows us to associate what was said with a truth-condition for it. So it does not point to a function, fixed by meaning, from speakings to thoughts expressed in them.

A second contrast between present phenomena and such things as 'I' and 'now,' traditionally conceived, is suggested by this remark of Frege's:

> the content of a sentence often goes beyond the thought expressed by it. But the opposite often happens too; the mere wording, which can be made permanent by writing or the gramophone, does not suffice for the expression of the thought.… If a time indication is conveyed by the present tense, one must know when the sentence was uttered in order to grasp the thought correctly. Therefore the time of utterance is part of the expression of the thought.… The case is the same with words like 'here' and 'there.' In all such cases the mere wording, as it can be preserved in writing, is not the complete expression of the thought; the knowledge of certain conditions accompanying the utterance, which are used as means of expressing the thought, is needed for us to grasp the thought correctly. Pointing the finger, hand gestures, glances may belong here too. (Frege, 1977, pp. 10–11)

We begin with the idea that sentences are related to thoughts in this way: for each sentence there is a thought which is the thought it expresses.[5] With indexicality, we lose that idea. There is no particular thought which is the one the sentence 'I am here' expresses. Perhaps, though, we may regain that idea if we permit ourselves to generalize the ordinary notion of a sentence. Ordinarily, we think of a sentence as a string of words. Suppose, though, we drop that idea. Let us call something a symbol if it has two features. First, it is individuated by purely non-semantic features, as a word might be individuated by its shape.[6] Second, it has semantic properties, where we will take that to be so if it makes a definite, specifiable semantic contribution to the whole, or wholes, of which it is a part. We might regard a (generalized) sentence as a structured set of symbols in this sense. So, if Frege is right about its semantic contribution, a time of utterance may be a symbol, and hence a constituent of a sentence in this sense. An utterance 'The leaves are green' in July would then count as a different sentence from an utterance, 'The leaves are green' in October – an odd, but coherent, way to speak.

If the only deviations from the rule that, for each sentence, there is the thought it expresses are represented by the sort of case Frege has in mind, then we may now regain the initial idea in this form: for each *generalized* sentence, there is a thought which is the thought it expresses. But the phenomena exhibited by (1) cannot be domesticated in this way. There is no identifiable feature of a speaking of (1) which counts as a symbol in the present sense, and whose semantic contribution to the speaking is identifiable with precisely

the set of truth-involving properties (1) would have so spoken. If the phenomena (1) exhibits are pervasive, then even a generalized sentence, no matter what extra symbols it contained, might be used to say any of many things.

Wittgenstein held that any symbol is open to different interpretations; and that under different circumstances, different identifications of its content would be correct. That is the moral of his discussion of rules and what they instruct (Wittgenstein, 1953, §§84–87). His arguments apply as well to generalized symbols as to others. If he is right, then the demonstration omitted here, that the parameter approach *cannot* work, is anyway to be found.

4 Implicature

> Suppose that I were the doctor and a patient came to me, showed me his hand and said: 'This thing that looks like a hand isn't just a superb imitation – it really is a hand' and went on to talk about his injury – should I really take this as a piece of information, even though a superfluous one? (Wittgenstein, 1969, §461)

> I am sitting with a philosopher in the garden; he says again and again, 'I know that that's a tree,' pointing to a tree that is near us. Someone else arrives and hears this, and I tell him: 'This fellow isn't insane. We are only doing philosophy.' (Wittgenstein, 1969, §467)

Wittgenstein cites some bizarre things to say. We do not say such things, barring very special occasion to do so. But what does that mean? Suppose one says them anyway. Despite the oddity, might one have spoken truth?

The philosopher does acrobatics recklessly close to the tree. 'That's a tree over there,' someone warns. 'I know that's a tree,' he replies testily. 'Well, then, shouldn't you be more careful?' Here the philosopher speaks truth. So, one might reason, he *does* know these things. But one cannot cease to know things, or so it seems, *just* by moving from one conversation to another. So however bizarre saying so may be in other cases, for all that, he speaks truth there too. So one might reason.

But this is a bad argument. For it *may* be that words like 'I know I'm wearing shoes' vary their semantics from speaking to speaking. If some speakings of them speak truth, that does not mean that all will. We cannot generally reason: Pia spoke truth when she called the leaves green; so if I call them green, I will speak truth too. That was the moral of §2. The point would be, not that the philosopher ceases to know something by changing conversations, but rather, that on one occasion he counts as knowing such-and-such, on another not.

There is, though, a form of account on which many bizarre things we 'would not say,' would, for all that, be true. The idea is due to H. P. Grice. The starting point is the observation that saying is only one of numerous ways for words, or speakers of them, to represent things as so. There is also implying, suggesting, insinuating, presupposing, and so on. *That* insight did not originate with Grice. Grice, though, concerned himself with a particular class of such representations, which he called implicatures, using the verb 'implicate' for the sort of representing in question. Implicatures come in two sorts: conventional and conversational. Conventional implicatures are features of the meanings of the terms involved. They are illustrated by 'Pia dissuaded Tod from leaving,' and 'Sam struggled to reach the lectern.' The first represents Tod as at least having thought of leaving; the second represents Sam as facing some obstacle to reaching the lectern. But the first does not *say* that Tod had thought of leaving, nor the second that there was an obstacle. That does not yet mean that,

for example, the second might be *true* were there no obstacle. It leaves it obscure what could make it so. But it may facilitate arguing the point. In any event, just as to use 'It's green' to mean what it does *is* to call something green, so to use 'struggle' to mean what it does, in a case like the above, is to suggest or imply that there is an obstacle. Grice suggests that it is difficult to produce words with a conventional implicature without implicating that. Such implicatures are not, or hardly, what Grice calls 'cancellable.' That he takes to be a main identifying feature of them.

Some implicatures, Grice notes, arise only on certain speakings of words, so *are* cancellable. These Grice calls conversational implicatures, and he explains them thus (though in much greater detail than given here). In normal conversation, we represent ourselves as observing certain maxims, and may be supposed to do so. Grice calls these *conversational maxims*. Examples are: be cooperative, be brief, be informative, and be relevant. Sometimes a speaker *seems* to violate some of these maxims. But it may be that he would not have if such-and-such, and it may be unreasonable to take the speaker to be violating them. We may then reason thus. The speaker said that P (in saying 'W'). Saying P (or saying it in 'W') would violate the maxims unless Q. The speaker was not violating the maxims. So (according to him) Q.

A speaker may intend for us to avail ourselves of some inference of this sort, to a given conclusion that (according to him) Q. It may be part of the proper understanding of his words that he so intends. In that case, the speaker has, or his words have, conversationally implicated that Q. For example, Pia may say, 'Jones submitted a sequence of English sentences, divided into paragraphs, and titled "What is truth?"' If this is merely a way of saying that Jones submitted an essay, then it violates the maxim of brevity. Pia would not do *that*. So, by the suggested sort of inference, we may conclude that there is, according to Pia, something which distinguishes Jones's work from a proper essay – perhaps its incoherence. It may have been given to be understood that we were so to reason. In that case, the conclusion was conversationally implicated.

The notion of conversational implicature points to a particular sort of understanding some words, on some speakings, may bear. Nothing in the pragmatic view suggests that there should not be such understandings. Note, though, that, as Grice insists, for Q to be conversationally implicated in words 'W,' Q must follow from what 'W' said, or the fact that 'W' said it, or both. So we might ask what Grice thinks words say. He is quite clear about that:

> In the sense in which I am using the word *say*, I intend what someone has said to be closely related to the conventional meaning of the words (the sentence) he has uttered. Suppose someone to have uttered the sentence *He is in the grip of a vice*.… One would know that he had said, about some particular male person or animal x, that at the time of the utterance … either (1) x was unable to rid himself of a certain kind of bad character trait or (2) some part of x's person was caught in a certain kind of tool or instrument … But for a full identification of what the speaker had said, one would need to know (a) the identity of x, (b) the time of utterance, and (c) the meaning on the particular occasion of utterance, of the phrase *in the grip of a vice* [a decision between (1) and (2)]. (Grice, 1989, p. 25)

This is just the rejected conception of saying. On it, for example, bracketing lexico-syntactic ambiguity, we can always form a guaranteed-true report, in indirect speech, of what was said in any arbitrary speaking of given words: if the words were 'The leaves are green,' then that the relevant leaves were, at the relevant time, green. To think that is to miss the possibility

of occasion-sensitivity in the content of 'green.' So Grice's conception of saying cannot be assumed in any argument directed against an instance of the pragmatic view.

Grice aimed to resuscitate views fallen into disrepute, largely through what were, in effect, early applications of the pragmatic view. For example, the idea of conversational implicature was first developed specifically in aid of reviving some notion of a sense datum. With that in mind, let us return to the bizarre remarks with which this section began. Consider 'I know that that's a tree.' It would usually be bizarre to say that, for example, where the tree was in plain view and no doubt of any kind had arisen as to whether it was a tree. Grice invites us to entertain the possibility that the reason we would not say such a thing in such circumstances is that if we did, we would conversationally implicate something not so. He means that idea to encourage us to ask whether what would be said if one did so speak is anyway something true, or rather something false; and to expect one choice or the other to be *correct*.

In using 'know' bizarrely we *may* conversationally implicate something (though there is a problem if conversationally implicating that Q absolutely requires saying that P). But the pragmatic view offers another explanation of why, in some situations, we would not say 'I know that…' Suppose that 'know' may make any of many distinct semantic contributions to wholes of which it is a part, and varies its contribution from one speaking to another. Then, describing someone as he is at a time, we would, on some occasions, say something true in saying him to know that X is a tree, and, on other occasions, say something false in saying *that*. For there are various things to be said in so describing him. In that case, circumstances of a speaking of 'N knows …' may confer on it a supplement to the content provided by the meanings of the terms alone. For some such supplements, the result will be stating truth; for others it will be stating falsehood. But *some* circumstances may fail to confer a supplement of either of these sorts. Words produced in such circumstances would have a content still supplementable in either way. But a content still so supplementable can require neither truth nor falsity. Speak, in those circumstances of N knowing that it's a tree, and one will fail both at saying what is true and at saying what is false. Nothing either so or not-so will have been said to be so. Recognizing that, where it is so, may make one refrain from so speaking. In that case, the idea, encouraged by Grice, that if we said it anyway we would at least say something true or else something false, is simply a mistake. In that case, conversational implicature could not be a consequence of the fact of having said *that* such and such. There is no such fact.

That the content of words is consistently supplementable in more than one way is not in itself a block to those words stating truth. It is so only where different such supplements, or different ones within some range of reasonable ones, yield different results as to truth – where, that is, the content to be supplemented is compatible both with truth and with falsity. So it just *might* be that if you say irrelevantly, pointing at your brogues, 'Those things are shoes,' there is no compelling reason to deny that you have spoken truth (though the situation changes if you are wearing four-eyelet low moccasin boots, or even just moccasins). That is typically not how it is for philosophically sensitive terms like 'know.' That is one lesson the long history of skepticism teaches us. (If there must be an *occasion-insensitive* answer, just when *does* someone count as knowing there is a tree before him?)

This last point shows the problem in applying the notion of implicature where it is meant to carry philosophic baggage, notably where it is meant as a way of dismissing claims about what 'we would not say' as philosophically irrelevant. Where those claims point to occasion-sensitivity they are philosophically highly relevant. It is all very well to insist, for example,

that either Sam does or doesn't now know that he is wearing shoes, full stop; and that if you said, bizarrely, 'Sam knows he is,' you would either state truth or state falsity. Sooner or later, though, one must choose. Which is it? If, applying the pragmatic view, we carefully assemble a perspicuous view of the *different* things we at least take ourselves to say to be so, on different occasions for speaking of Sam, in saying him to know precisely that, then *either* there is a principled way of choosing between them (or choosing a further candidate) by appealing to what is recognizably so about what 'know' means, *or* they show that no one answer to the question is the right one occasion-independently. Prospects for the first alternative are dim.

5 Metaphysics

The English 'is green' speaks of a certain way for things to be: green. One might say that it speaks of a certain property: (being) green. If we do say that, we must also say this about that property: what sometimes counts as a thing's having it sometimes does not, so that there are, or may be, things which, on some occasions for judging, count as having the property, and on others do not. If for a property to have an extension (at a time) is for there to be a definite set of things (at that time) which are just those things (then) with that property, then this property does not have an extension, even at a time. Better put, it makes no sense to speak of 'its extension.'

Is all this just vagaries of the English 'is green'? Two related questions arise. First, might there be predicates which did not vary their contributions to what was said with them in the way that 'is green' does? If we said such a predicate to speak of a property, that property *would* have an extension, at least at a time. Such a predicate could not vary its contributions to wholes so that, in ascribing that property to an object (at a time) it would be possible to speak truth and also possible to speak falsehood. So there would be no call for saying of anything that it sometimes counted, and sometimes didn't, as having (at a given time) that property. Second, can we preserve the idea that (genuine) properties have extensions by supposing that predicates like 'is green' simply refer to different properties on different occasions (and that it is by their thus varying their referent that they make different contributions to different wholes)?

Why might one want properties to have extensions? First, one might think that we can gain this for properties by definition – by 'property' we just mean what has an extension – and that extensions are convenient means for counting properties (as one or two). Second, one might take such a view of properties as mere sane realism. We cannot change, say, the way a cow is by thinking about it. As a rule, the cow stays just as it is no matter how we think of it. And we may read, or misread, that sane thought thus: those ways for things to be which are, or count as, ways the cow is count as ways the cow is no matter how we think about the cow, or them. So for any genuine way for things to be, either the cow is that way (at a time), or it is not, *punkt*. The same goes for any other object. In which case, genuine ways for things to be have extensions (at times). But whatever there is in favor of this line of thought, I suggest that both our questions merit negative answers.

I begin with the first. I will state the main point, though there is here no space for detailed argument. Once we fix what 'is green' speaks of – green – we then note that there are different possible understandings of what it would be for an object (or some objects) to be *that* way (green). These are possible understandings in that they represent what one *might*

regard as a thing's being green. So, for each, some item may be said, in calling it green, to be green on *that* understanding of its being so. And for each, that may be the *right* understanding (on some occasion) of what being green would come to. 'Is green' provides a particular description for things, expresses a certain concept. What is said in using it depends not only on what that description is, but on how that description, or that concept, is, or would be, applied in fitting it to particular circumstances of its use.

Suppose, now, that we identify an understanding of being green – say, the understanding on which an item was said to be green in some particular speaking of 'is green.' We now introduce a predicate – say, 'is green*' – which, by stipulation, is to mean *is green on that understanding of being green*. This predicate speaks, as it were, of a finer-grained property than 'is green' (as such) does. May *this* predicate make different contributions to what is said in wholes of which it is part? It may if there are different possible understandings of what it would be to be green on that understanding; two different things to be said as to whether such-and-such *is* being green on that understanding of what it would be to be so. As far as we can tell, this always will be so. We understand, for example, that paint is to count as changing color, and not as hiding it. We now encounter a rather poor paint job: you *could* say that it covered the original color, but you could view the original color as still showing through enough that the object had not yet been made the color of the paint, even on the indicated understanding of its being that color. An understanding of being green, in so far as we can identify one, seems unable to foreclose in principle on the possibility of differing but, apart from particular surroundings, equally sane and sensible views of what *that* understanding entails.

A predicate about which the pragmatic view was wrong would be one which did not admit of different possible understandings of what it would be for some item to fit the description which that predicate provides (or for the description to fit some item). The right understanding of it would foresee every eventuality in or to which the description might be applied. There is reason to think that no such predicate is available to human beings, at least given the way we in fact cognitively conduct our affairs. Again, what is said in applying a given description depends on *how* it is applied, and how, in given circumstances, it ought to be.

Now for the second question. First, if the first point is correct, then no understanding *we* could have of being green, so none that might attach to a particular use of 'is green,' would be one on which 'is green' spoke of a property, if a property must have an extension. To paraphrase Wittgenstein, we refine our concepts, or understandings, for particular purposes – so that *in fact*, in the situations we face or expect, unclarity as to what to do or say does not arise. In doing that we neither reach, nor aim at, that absolute clarity on which we would speak of what had definite extensions. Where 'is green' has made different contributions to different wholes, we may identify different things for it to have spoken of each time – being green on this understanding, and being green on that one. So we may see the predicate as varying its reference across speakings of it. But we must not mistake these different things for properties with extensions. Second, if we cannot have a predicate for which the pragmatic view does not hold, then, equally, we have no means for specifying properties to which extensions may sensibly be ascribed. In any event, the phenomenon we have to deal with is not merely that predicates vary their contributions to wholes, but also that, whatever a predicate may be said to speak of – being such-and-such – what would sometimes count as an item's being *that* other times would not.

6 Perspective

Given words may have any of many semantics, compatibly with what they mean. Words in fact vary their semantics from one speaking of them to another. In that case, their semantics on a given speaking cannot be fixed simply by what they mean. The circumstances of that speaking, the way it was done, must contribute substantially to that fixing. As pointed out earlier, this does not mean that there is a function from certain parameters of speakings to semantics, taking as value for each argument the semantics words would have where those values held. It thus also does not mean that there might be a precise theory, generating, for each semantics words might have, necessary and sufficient conditions for their having that. Still, we may describe how circumstances do their work.

Here is one thought. The words 'is green' are a means which English provides for calling things green (describing them as green, etc.). If, in speaking English, you want to call an item green, those words will do. Speak them literally, seriously, and so forth, and you will then count as having done just that. The truth of what you say in calling an item green should turn precisely on whether the way that item is then counts as its being green. These two remarks jointly identify which truth-involving properties any such words must have: they are true of, and only of, those ways for things to be which counted, at their speaking, as the item they spoke of being green. Similarly for other English predicates.

Where you called an item green, the truth of your remark turns on whether it *then* counted as being green. On different occasions, different ways for an item to be would count as its being green. That variation means that, on different occasions, calling an item green will confer different truth-involving properties on your words. Consider two occasions which differ in this respect. On each, words which call an item green will have some set of truth-involving properties, which is, therefore, a possible set of such properties for words with that content to have. Each such set of truth-involving properties, and the property of calling that item green, cohere on at least some occasions for so describing things. But those truth-involving properties *cannot* be those of words with that content produced on the other. That would not correspond to what, on the other, counts as something's being green. So each of the above semantics, available as it is on some occasions, is unavailable on others. I can sometimes speak truth in calling painted leaves green; but I cannot do so in circumstances where their being so painted does not count as their being green.

Let us pursue this thought. Consider:

(2) Today is a sunny day.

Spoken on day D, (2) would, typically, speak of day D. It would also identify the day it speaks of in a particular way: it speaks of that day as the day of its speaking, and represents it as identified by that fact. Since some speaking of (2) has both the semantic properties just mentioned, the two jointly form a semantics which is at least sometimes coherent. Let D* be the day after D. Words produced on D* could not have the semantics just mentioned. They could not speak of D and say it to be sunny while, on their proper understanding, identifying the day they speak of as the day of their speaking. On day D, we may express, or think, a thought with both those features. On other days (in normal circumstances) we cannot. Let us say that words with a semantics which is only sometimes available, in the above sense, express a perspectival thought, and have a perspectival content.

Now the point of the discussion of 'is green' may be put this way. Perspectival thought is the normal and pervasive case. On one occasion, we call an item green (at a time), and thereby produce words with such-and-such truth-involving properties. On another occasion, we may, if we like, say the same item to be green (at that same time). But our doing that may require that our words have quite different truth-involving properties. Those of our first remark may not correspond to what would count, on the occasion of this further speaking, as that item's being green. If that is right, it is fair to suppose that perspectival thoughts are the typical sort of thoughts we think. One might say: we relate cognitively to the world in essentially perspectival ways.

Now consider two minor puzzles. First, I have said there is something true, and also something false, to be said of given leaves, and their condition at a given time, in saying them to be green. How can this be? Consider the true thing to be said. What could make it true, other than the fact that the leaves are green? But, if that is a fact, how could one speak falsehood in saying no more nor less than that about them? Second, if there *are* those two things to be said, then *say* them, or rather, state the true one and deny the false one. To do so, you would have to call the leaves green, and then deny that they are that, as in 'The leaves are green, and the leaves are not green.' But that is a contradiction, so cannot be true. So what the pragmatic view requires that it be true to say is something it could not be true to say. So the view is wrong.

The first puzzle's rhetorical question has a non-rhetorical answer. What could make given words 'The leaves are green' true, other than the presumed 'fact that the leaves are green,' is the fact that the leaves *counted* as green on the occasion of that speaking. Since what sometimes counts as green may sometimes not, there may still be something to make other words 'The leaves are green' false, namely, that on the occasion of *their* speaking, those leaves (at that time) did not count as green.

As for the second puzzle, we are challenged to say something literally unsayable – *not*: sayable-but-false, but rather not sayable at all. We ought to decline the challenge. On some occasion, words which call given leaves (at a time) green may (thereby) have truth-involving properties in virtue of which they are true. On some other occasion, words which deny those same leaves to be green may similarly be true. But given the way (described above) in which occasions work to forge a link between content-fixing properties and truth-involving ones, there is no occasion on which both these feats could be accomplished at once; so none on which 'The leaves are green and the leaves are not green' could have the semantics which a conjunction of those two truths would have to have. If the occasion is one on which the way those leaves are counts as their being green, then no words could have the semantics of the true denial; and *mutatis mutandis* if on the occasion the way the leaves are does not count as their being green. Each of the thoughts provided for above is a *perspectival* thought; and, in virtue of its perspectival character, unavailable to be expressed at all on any occasion on which the other is expressible.[7] The nature of semantic variation thus allows us to decline the challenge.

These are banal examples. In philosophy, neglect of perspectival thought often leads to more excitement. A philosopher may sense, for example, that our concepts apply as they do against a background of our natural reactions; if we naturally viewed things *quite* differently, we might apply the concepts we now have so as to speak truth in saying what it would not now be true to say. Asked to express some such truths, the philosopher is reduced to nonsense. Naturally enough. He was describing other perspectives. Some things said truly from them are not so much as expressible at all from his own.

7 Thoughts

Frege writes,

> Without offering this as a definition, I mean by 'a thought' something for which the question of truth can arise at all. (Frege, 1977, p. 4)

Thoughts, for Frege, are not words. For him words are true only in a derivative sense: just in case they express a thought which is. For words are always open to, and in need of, interpretation. They are true, if at all, only on a given understanding of them (even if it is their proper understanding). Words 'Mary had a little lamb' may be a remark on husbandry, or one on gastronomy and, perhaps, true if understood the first way, false if understood in the second. Truth and falsity seem to correspond to understandings words may have, rather than to the words themselves (which Frege conceives as a quite different matter). It is the understandings, as opposed to the words, which settle questions of truth and falsity. So, on his view, it is for understandings, and not for words, that questions of truth and falsity arise. Words, apart from an understanding, could not be true or false at all.

If words admit of interpretations, then conceivably they may bear different understandings on different occasions for understanding them. Such shifts in interpretation could bring with them shifts in truth-value. So if words were the primary objects for which questions of truth arose, it would be conceivable, for any sort of semantic object, that one and the same item should count as true on one occasion for assessing it, false on another.

Thoughts, for which questions of truth are, strictly speaking, to arise, are meant to be free in principle of both of the above features. They are to be absolutely immune to interpretation; and they are to be true or false absolutely, independent of the ways, if any, in which they enter into our thinking. On Frege's view, only such semantic objects could be material for logic.

We may extend the notion of semantic property so that thoughts have a semantics too. The semantic features of a thought will be just those features by which one thought may be distinguished from another. Among these will be such things as being about eating ovine, and such things as being true if Mary ate a bit of ovine, hence, on the above plan, both truth-involving and content-fixing properties. Its truth-involving properties are meant to be just those its content requires. Moreover, it is meant to have all this semantics intrinsically: any thought, no matter how encountered, is that thought iff it has that semantics. This means that the content of a thought – unlike the content of words – must determine its truth-involving properties *inexorably* (to coin a term): there are no two sets of truth-involving properties such that an item with that content might have the one but not the other, and also vice versa; there is *one* set of truth-involving properties which is *the* set any item with that content *must* have. For if not, then a thought's having that content might, on some occasions, make it count as having one set of truth-involving properties, and on others make it count as having another, counter to the tenet that every thought has its truth-involving properties intrinsically.

Why must thoughts have inexorable content? Suppose C is a non-inexorable content. Then there might be an item with C and truth-involving properties T, and an item with C and distinct truth-involving properties T*. But truth-involving properties are meant to be those which content requires. So these must be two items differing in further content-fixing features. This means that an item with C is, so far, open to interpretation: it might, for all

that, bear any of several distinct understandings. That is to say: it might, for all that, be, or (if words) express, or represent, any of several distinct thoughts. So C is not the (whole) content of a thought.

Thoughts are identified precisely by their semantics, whereas words are identified by shape, syntax, or spelling, or by the event of their production. The identity of words leaves their content open. So the content of given words must depend on further factors: on the character of their surroundings. This leaves it open that their surroundings might, on some occasions of considering them, count as conferring one semantics on the words, while on other such occasions those surroundings might count as conferring another. In that way, the semantics of words – how they are rightly understood – may be an occasion-sensitive affair. By contrast, the semantics of a given thought is meant to depend on *nothing*. So there are no such possibilities for variation across occasions in the semantics a given thought *counts* as having.

Thoughts, as thus conceived, are not open to interpretation. They are what Wittgenstein called 'shadows': semantic items interpolated between words and the states of affairs that make words true or false, and somehow more closely tied to those states of affairs than mere words could be. About shadows, Wittgenstein said:

> Even if there were such a shadow it would not bring us any nearer the fact, since it would be susceptible of different interpretations just as the expression is.[8]

How could this be true of thoughts? Could thoughts admit of interpretation? If so, how?

There are too many strands in our inherited notion of a thought to unravel them here. But here is a sketch of a framework for relevant issues. To begin, one *might* think to buy the semantic absoluteness of a thought – its immunity to interpretation – by stipulation. Wherever I would say something to be so in saying 'S,' and it is determinate what, I may, it seems, refer to a thought in saying 'the thought that S.' I may also say, correctly, it seems: 'The thought that S is true iff S.' In saying that, I ascribe a set of truth-involving properties to the thought I refer to; in fact, whatever such properties my words 'S' then had. For I say the thought to be true exactly where what is so according to my words 'S' is so. So, it seems, we might stipulate that the thought I thus refer to is precisely the one with those truth-involving properties.

This is not quite enough. A thought cannot *just* have truth-involving properties. It must have a content. What content should that be? Here we come up against another strand in the conception of a thought. A thought is meant to be something that can be expressed in various words, or speakings, on various occasions. If you now express a thought, I can later express that very thought virtually whenever I like. On any plausible version of that view, words W and W* may express the same thought while differing in content. Frege gives this example:

> If someone wants to say today what he expressed yesterday using the word 'today,' he will replace this word with 'yesterday.' Although the thought is the same, its verbal expression must be different in order that the change of sense which would otherwise be effected by the differing times of utterance may be cancelled out. (Frege, 1977, p. 10)

The word 'today' brings with it a different contribution to content than the word 'yesterday.' Frege's two sentences are not alike in content-fixing properties. Yet, for good reason, Frege

takes it that the one sentence, produced under certain circumstances, would express the same thought as the other sentence produced under certain others. If so, then the content-fixing properties of that thought are liable to vary across occasions.

The question is: just *how* may content vary while words express the same thought? One idea would be that W and W* express the same thought only if they apply the same concepts to the same objects. But this will not do. It does not even allow for Frege's example. It collapses completely if we return to the notion of perspective. On some occasions, in calling given leaves green one would state truth; on others, in calling those leaves green one would state falsehood (and not because the leaves changed). Apply a given concept to the leaves in different surroundings and you will produce words with very different truth-involving properties. The semantics of some such words, produced in given surroundings, is unavailable in other surroundings for any words. Words with the *content* of those words, in the other surroundings, may have truth-involving properties so different that, at least for some purposes, we cannot take them to have expressed the same thought. The false remark about the leaves, for example, was not the same thought as the true remark. So if, in the changed surroundings, one wants to express the same thought again, one must *not* speak of the same concepts and objects. What it would take to express the same thought again is nothing more nor less than an adequate paraphrase. If the original words were 'The leaves are green,' then, depending on surroundings, an adequate paraphrase might be 'The leaves are painted green.'

There is no space here for an account of what makes paraphrases adequate. But here are two remarks. First, adequate paraphrases may need to share crucial or relevant truth-involving properties; but they are unlikely to share *all* truth-involving properties. In remote enough circumstances, leaves may be green in the sense in which they were said to be in a given 'The leaves are green,' but not painted green (perhaps dyed); though, for current purposes, 'The leaves are painted green' was an adequate paraphrase. Second, suppose on an occasion I express a thought in saying 'The leaves are green.' Then whether, on another occasion, words W are an adequate paraphrase of what I said may well depend on the occasion for the paraphrase, and perhaps, too, on the occasion for considering that occasion.

Thoughts viewed from this position lose their claims to have some *one* semantics intrinsically, and to be immune to interpretation. If, with perspective in mind, we ask what would count as producing some given thought again, and if we consider all the occasions for posing that question, we see how *that* thought may count on some occasions as having semantics which it would not count as having on others. For it may on some occasions admit of paraphrases it does not admit of on others. Nor need it ever have an inexorable content. To see how thoughts admit of interpretation, one need only know how to look for occasions for interpreting them.

8 Concluding Remarks

There is much left to discuss, but no space left to discuss it. It is thus time to commend the subject to the reader. The pragmatic view gives a substantially different form to virtually every philosophic problem, not just in philosophy of language, but wherever puzzles arise. The new form may make some of these problems more tractable. For a start we will need new conceptions of logical form, and of such related notions as intensionality. These may yield new things to say on such questions as whether 'if–then' is transitive. We may then

take a fresh look at what we say of people in ascribing propositional attitudes to them, and at understanding itself. Such a look, I predict, would make philosophy of psychology take a fresh course. It is also worth a look, from the pragmatic view, at problems of knowledge, of explanation, of freedom and responsibility, and so on. Some of this work is begun. There is much left to explore.

Notes

1 For some more discussion see my (1989, especially ch. 1).
2 Throughout I leave lexico-syntactic ambiguity aside.
3 I modify Frege's example slightly. His discussion is in Frege (1977, p. 27).
4 See my (1991) for further discussion.
5 Once again, ignore lexico-syntactic ambiguity.
6 Strictly speaking, this is false of words (consider, e.g., homonyms). But ignore that for now.
7 More precisely, any occasion on which a thought with the semantics of the first is expressible is *ipso facto* one on which a thought with the semantics of the second is not. I do not mean to prejudge questions of thought-identity.
8 Reported by Moore (1954/1993).

References

Frege, G. 1977. "The thought." In *Logical Investigations*. Oxford: Blackwell.
Grice, H. P. 1989. "Logic and conversation." In *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Kalish, D. 1967. "Semantics." In *The Encyclopedia of Philosophy*. New York: Macmillan.
Lewis, D. 1972. "General semantics." In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 169–218. Dordrecht, Netherlands: Reidel.
Moore, G. E. 1954. "Wittgenstein's lectures in 1930–33." *Mind*, 63(249): 1–15. Reprinted in *Philosophical Occasions 1912–1951*, edited by J. C. Klagge and A. Nordmann, p. 59. Indianapolis, IN, and Cambridge: Hackett, 1993.
Platts, M. 1980. "Introduction." In *Reference, Truth and Reality*. London: Routledge and Kegan Paul.
Travis, C. 1989. *The Uses of Sense*. Oxford: Oxford University Press.
Travis, C. 1991. "Annals of analysis." *Mind*, 100(398): 237–264.
Wittgenstein, L. 1953. *Philosophical Investigations*. New York: Macmillan.
Wittgenstein, L. 1969. *On Certainty*. Oxford: Blackwell.

Further Reading

Austin, J. L. 1961a. "Truth." In *Philosophical Papers*. Oxford: Oxford University Press.
Austin, J. L. 1961b. "How to talk." In *Philosophical Papers*.
Austin, J. L. 1961c. "Other minds." In *Philosophical Papers*.
Austin, J. L. 1962a. *How to Do Things with Words*. Oxford: Clarendon Press.
Austin, J. L. 1962b. *Sense and Sensibilia*. Oxford: Oxford University Press.
Barwise, J., and J. Perry. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.
Cartwright, R. 1987. "Propositions" and "Propositions again." In *Philosophical Essays*. Cambridge, MA: MIT Press.
Dummett, M. 1993. "Mood, force and convention." In *The Seas of Language*. Oxford: Oxford University Press.

Fauconnier, G. 1985. *Mental Spaces*. Cambridge, MA: MIT Press.
Grice, H. P. 1975. "Logic and conversation." In *Syntax and Semantics*, vol. 3, edited by P. Cole and J. Morgan. London: Academic Press. Reprinted in *Studies in the Way of Words*. Cambridge, MA: Harvard University Press, 1989, pp. 22–40.
Grice, H. P. 1978. "Further notes on logic and conversation." In *Syntax and Semantics*, vol. 9, edited by P. Cole, pp. 113–127. London: Academic Press. Reprinted in *Studies in the Way of Words*, pp. 41–57.
Grice, H. P. 1989. "Retrospective epilogue." In *Studies in the Way of Words*.
Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–561. Oxford: Oxford University Press.
Sperber, D., and D. Wilson. 1986. *Relevance*. Oxford: Blackwell.
Stalnaker, R. 1972. "Pragmatics." In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 380–397. Dordrecht, Netherlands: Reidel.
Strawson, P. F. 1952. *Introduction to Logical Theory*. London: Methuen.
Travis, C. 1996. "Meaning's role in truth." *Mind*, 105(419): 451–466.
Ziff, P. 1972a. "Understanding." In *Understanding Understanding*. Ithaca, NY: Cornell University Press.
Ziff, P. 1972b. "What is said." In *Understanding Understanding*.

Postscript

CHARLES TRAVIS

In Retrospect

Eighteen years ago I set out, above, a view on the relation of language to what is said in using it. I still think the view right. What *has* changed in the interim is the case I would make for it. Eighteen years ago this case consisted, in essence, in presenting the view and then disarming various anticipated objections to it. In the interim my ambitions have grown. Having stated the view, I would now aim to identify features of thought and its expression which would dictate occasion-sensitivity, both of words and of what they speak of. Not whether occasion-sensitivity is coherent, or *possible*, but rather why our form of thought requires it.

There are several lines one might pursue in making such a case. In present confines I merely outline one. It starts from an idea about what words are for. It continues with another about how deeply thought is world-involving. To arrive at the first we can start from Frege's idea of a *thought* (*Gedanke*). In a thought things are represented as some way there is for them to be. The thought itself may be said so to represent things. The thought is an abstraction from historical instances of representing (something) as (something): instances of thought-expression. It does not represent in the same aspect of the verb as that in which *we* represent in saying things. It is more an aspect in which for the *thought* to represent as it does is for that to be how *one* would represent in its expression. The thought abstracts from activities, or stances, in order to serve a particular purpose. Thoughts are to be that which laws of logic govern. Logic (for Frege) answers the question how one must think to reach the goal truth – but only in so far as the answer to that question is given by what being true is as such (1897, p. 139). The laws of truth unfold the concept *being true* (1918, p. 59). A thought identifies a particular way for *how* things are represented being to make truth depend on *what* is so represented. Thus (e.g.) a thought cannot be an object of sensory awareness, nor have any identifying feature which is (cf. Frege, 1918, p. 61). How

representing looks or sounds has nothing to do with when the way things are thus represented as being is a way things are. There is no such thing as two thoughts – a true and a false one, say – distinguished from each other in that one is written in Gothic script, the other not.

Such already tells us something Frege often stressed: a thought is quite un-language-like. For example, it is absolutely insusceptible to the kind of structure – syntactic – by which a *sentence* of some language is identified as the one it is. A thought and a sentence cannot share a structure. First, a sentence is assigned a structure in its generation by the syntax of its language. To be that sentence is to be *so* structured, full stop. Whereas for thoughts, following Frege, the whole thought comes first. It is, of course, decomposable into elements. But then in many different ways, and – unlike a sentence – into any of many diverse sets of elements. No one decomposition, as Frege tells us, can claim objective priority. (*Vide* Frege 1882, p. 118; 1892, pp. 199–200; 1919, p. 273.) Second, a thought is structured (or structurable) in terms of relations defined in terms of being true, whereas a sentence is structured in terms of relations defined in terms of being a sentence in its language. A structure of a thought on a decomposition is not a syntactic structure. (*Vide*, e.g., Frege 1897, p. 154.)

A thought, as noted, is invisible. Why should a sentence need to have a visible, or audible, form? Here is a clue. Pigs, or sheep, have a particular visible form, here a look. There is that distinctive porcine look; similarly that ovine look. A pig is (often) thus recognizable by sight. You can tell a pig by its porcine look. Perhaps, in some similar way, a sentence's look – that 'Pigs fly' look, so to speak, helps to make something recognizable. What? A sentence is not just all looks. It is (as a rule) also meaningful. What it means, or speaks of, matters in a specific identifiable way to what would be said in speaking it (as meaning what it does). So if you have recognized a sentence, in some production of it, by its looks and you know what it means (I here bracket ambiguity), and if that production was, recognizably, in an act of saying something, you have thus taken a significant step towards recognizing what was thereby said.

Saying something is a particular case of representing-as. It is (an act of) *authored* representing. It is an act *one* might be aware of. For things, in it, to be represented as a certain way is for such to be recognizable to *one*. A sentence is well designed to help in such an enterprise. Its perceivable form makes it identifiably instanced, occurring, independent of how, if at all, things are thus represented as being. In meaning what it does, it makes a specific contribution to identifying how things were said to be (or at least spoken of as being) in so employing it. Allorepresenting would not be impossible without this. But, arguably, without it we could hardly achieve the complexity and subtlety of the representing-as we in fact produce.

That language has the task of achieving recognizability, and that each meaningful expression of a language serves this in its own proprietary way, are not anywhere in serious dispute. The core idea I defended 18 years ago is what I would now put thus: serving the end of recognizability is the *sole* aim which linguistic meaning is tasked with serving; for an expression to mean what it does just is for it to contribute as it does to making the authored representing-as done in it recognizable as representing as it does. A language is *not* charged with itself containing truths and falsehoods – sentences which, just in meaning what they do, say what is either true or false (of things as they are) – nor, equally, with containing truths or falsehoods *of* n-tuples of given sorts – e.g., triples of an object, time, and place. I call this view the 'authoring tools view,' ATV for short.

Frege distinguished matters of being true from those of holding true – more generally, the psychological (in some broad sense) from the logical (again, in some broad sense). Following Frege we might distinguish between matters of holding forth as true – part of the psychological in this sense – and matters of being true. The business of linguistic meaning, the idea is, is matters of the first sort, not the second. Thus the contrast between sentences and thoughts, and the sort of structuring to which each is susceptible. The next step along this line of thought is to show that the means by which words achieve the recognition that they do (where to be taken as used as meaning what they do) are such as always to make room for occasion-sensitivity: the recognition a predicate would thus achieve of how a thing (or n-tuple) was said to be in using it to speak of what it does is always liable to vary from one such use to another. (For further discussion see my forthcoming a and b.)

The sentence 'Champagne bubbles' may be said, correctly, to represent champagne as bubbly. Is this not representing truly or falsely according as champagne is bubbly or not? But think of the aspect of the verb 'represent' here. Someone is asked whether champagne bubbles. He replies: "Well anyway the sentence 'Champagne bubbles' so represents it." But the sentence "Champagne does not bubble" represents it as not such. The one is no more reliable than the other. Neither lends any support to either view. Neither itself represents either truly or falsely. Similarly for 'say' and 'speak of.' To say a sentence to speak of thus and so is (roughly) to say: you would speak of that in using it on an occasion. Such does not require there to be just one way one could say things to be in speaking of them as *that*.

A sensory form would block a thought from filling its role in the phenomenon of being true. Equally, expressing as such truth or falsehood would block a sentence from filling its role in achieving recognition for holding forth as true. How to see this? One might first take a leaf from David Kaplan (*vide* Kaplan, 1989), then one from Hilary Putnam (*vide* Putnam, 1975a and 1975b). Kaplan agrees that for words to mean what they do is for them to contribute as they do to achieving recognition of the representing authored in their use. For a sentence to mean what it does is thus, for him, for it to have a 'character.' He departs from the present view in supposing its character to determine what would be said in it on any use (at least of some relevant n-tuple). On the present view, what is said, or even what is predicated, is *always* the upshot of *substantial* cooperation between meaning and circumstance. Still, he provides grist for our present mill.

A centerpiece in Kaplan's application of his distinction concerns the expression of singular thoughts – thoughts of some given object that *it* is thus and so. His aim is to defend a thesis about them: that for a given object and way for it to be, there is just *one* thought of that object that it is that way. Be *that* as it may. Anyway, singular thoughts *can* be expressed in using definite descriptions – a *prima facie* problem for his thesis. Kaplan uses character to disarm this; thereby a more general form of objection. At the reception, Alf spots Sid across the room at the buffet. Seeing what Sid (appears to be) doing, Alf remarks to his companions, 'That man who just fed the last slice of *foie gras* to his pet marmoset paints barns.' Alf thus expresses a thought, of Sid, that he paints barns. The description by which, in part, he succeeds in this identifies a person, Sid, in speaking of various ways for a thing to be: being the last slice of *foie gras*, being a pet marmoset, being one who just fed the former to the latter, being a man. It does not follow from this, Kaplan stresses, that the thought expressed does its representing in identifying anything by its being any of those ways for a thing to be. The description makes the thought expressed identifiable as the one then expressed. But it need contribute nothing to this beyond making some object (here Sid) identifiable as the one that thought is about. Which point is underlined by the fact that that object, Sid,

need not *be* as described in that description. Perhaps it is not the last slice. Perhaps it is not a marmoset but a rhesus. Perhaps the slice was not fed to it, but deftly pocketed in Sid's handkerchief pocket. And so on. What is needed for the description to achieve the recognition it is aimed at is no more than (roughly) this: that there should be an object who the speaker would then be (to be) understood to be supposing to be as thus described.

Two things need noting here. The first is the role of supposition (or close kin) in effecting the recognizability to be effected. The words Alf spoke do not speak of anyone in English. As thus spoken they purport to do so. The one this is (if any) is to be that one, if any, of whom, in the circumstances, the speaker would be to be understood to be supposing to be the one that description fit. The second is the conflict here between the demands of achieving recognition and those on contributing to a definite question of truth – to some determinate way for truth to turn on how things are. It would not have been much of a contribution to recognizability had Alf simply said, 'That object,' nor even (in a room replete with men) 'That man.' A richer understanding is called for as to *what* object. Thus the description, exploiting a particular striking bit of history which the object named *appeared* to have. The way a thing is thus described as being *is* an element in making recognizable what thought was thus expressed. By contrast, once a *thought* has been decomposed into an element which makes truth turn on how *Sid* is, and one which makes it turn on which objects paint barns, it is then otiose to seek further elements which make truth turn on further factors. We have already identified a way for truth outright to turn on how things are. Whatever other elements the thought had, they could not contribute to that. Which means, given what a thought is, that they could not be its elements. The means language provides for achieving recognition are thus the *wrong* means for constituting thoughts.

The next step is to generalize the role of supposition in achieving recognition so as to include using a *predicate* of a language to achieve recognition of what the thought expressed *predicates* of some object(s). What is the role of 'paints barns,' for example, in achieving recognition of how Alf, in those words, represented Sid as being? Here we need a leaf from Hilary Putnam's book. For many years now Putnam has shown us that the only ways we have of standing towards a way for things to be – of making it such-and-such that we were thinking, or speaking, of in thinking or speaking of things being thus and so – are intrinsically such that what it *is* we in fact thus speak of is liable to be something which is *not* what we had assumed, or even stipulated, it would be. We speak of, say, being the shortest path from one point to another. What we thus speak of is susceptible to proving something we *had* thought a shortest path could not be. Putnam shows that those thoughts available to us to think are deeply world-involving. Of course there would have been no thoughts about champagne had there been no champagne to think about. But moreover, if there is such a thing as being champagne, then what it would *be* for something to count as being that way – what would count as such – also depends on the course the world in fact has taken.

So the rule is: where we *do* speak of some way for things to be, the way we thus speak of is that way for things to be of which what would have been to be supposed, in our circumstances, as to what way this is would then have been to be supposed. If being gold would have been taken to be being a metal which was heavy yellow and malleable, then what that would have been to be supposed of is being a *metal* which, in pure state, is not yellow. Putnam's principal application of this idea was always to issues as to what words, for example, 'is yellow,' speak of as such; to what way for things to be being yellow is full stop. The aim is to generalize 'Kaplan's point' to predicates by generalizing Putnam's point to *acts* of using words to represent.

At yet another reception, Alf said to his companions, 'Sid paints barns.' He thus used 'paints barns' as meaning what it does. He thereby spoke of (someone as) painting barns – in the 'depict' rather than the 'cover' sense of 'paint.' What way for things to be would he then be speaking of? But I have already given this answer: being a barn painter. What remains to say? Many questions might remain as to which way for a thing to be being a barn painter is. Is it, for example, a way someone might be if only a weekend painter, or if he only paints converted barns (chic retreats)? Or ruined barns? Or only the stalls? Notably, is it a way such that Sid's being as he is would be a case of someone being it? Putnam's point now generalizes. The way one would then speak of in speaking of someone as a barn painter is that way of which what was to be understood to be supposed as to what way Sid was thus speaking of *was* then to be supposed correctly. Now the main point. What one would be understood to be supposing on one occasion as to what he was speaking of in speaking of being a barn painter need not be what one would be understood thus to be supposing on another. Correlatively, that of which he *did* thus speak is liable to differ from one such occasion to another.

Two morals. First, what one speaks of in *then* speaking of being a barn painter is what one would then be understood thus to speak of: that which one would then be speaking of in speaking of what being a barn painter was then to be understood to be. One thus speaks of being a barn painter as this is to be understood on the occasion. Which leaves room for how one thereby says someone to be to vary from occasion to occasion for so speaking. Second, the route from what was then to be understood as to what was thus being spoken of to what *was* thus being spoken of – and thus (via this route) from linguistic meaning to what *was* thus spoken of – is indirect enough for the service of the end of recognition to conflict with the service of that of being true. Which mandates ATV.

Such is a sketch. No room here for proof. It sketches one reason why ATV, with it occasion-sensitivity, *needs* to be the way our thought is structured. I will mention *one* other line. It focuses on a special case of representing-as: representing to be. In authoring such representing one assumes responsibility. If Alf's claim about Sid (a barn-inspired abstractionist) leads Pia to misestimate Sid entirely, she may or may not have right to blame Alf, depending on what it is Alf said in calling Sid a barn painter. For what *is* Alf to be held responsible? What responsibilities would he reasonably incur in and by *that specific act*? Well, how might that act reasonably be supposed to be contributing to pursuit, on Pia's part, of the thing for her to do? Such are questions whose answers would quite plausibly vary from act to act of calling someone a barn painter, according to the occasion there then was for doing so. Which would again yield occasion-sensitivity. But such is an idea, not even yet a sketch.

References

Frege, G. 1980 (1882). Letter to Marty, *Gottlob Freges Briefwechsel met D. Hilbert, E. Husserl, B. Russell, sowie ausgewählte Einzelbrefe Freges*. Hamburg: Felix Meiner.

Frege, G. 1892. "Über Begriff und Gegenstand." *Vierteljahrsschrift für wissenschaftliche Philosophie*, 16: 192–205.

Frege, G. 1983 (1897). "Logik." In *Nachgelassene Schriften*, 2nd edn, pp. 137–163. Hamburg: Felix Meiner.

Frege, G. 1918. "Der Gedanke." *Beiträge zur Philosophie des deutschen Idealismus*, 1(2): 58–77.

Frege, G. 1919. "Aufzeichnungen für Ludwig Darmstaedter." In *Nachgelassene Schriften*, 273–277.

Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–561. Oxford: Oxford University Press.

Putnam, H. 1975a. "It ain't necessarily so." In *Mathematics, Matter and Method* (Collected Papers vol. 1), pp. 237–249. Cambridge: Cambridge University Press.

Putnam, H. 1975b. "The analytic and the synthetic." In *Mind, Language and Reality* (Collected Papers vol. 2), pp. 33–69. Cambridge: Cambridge University Press.

Travis, C. Forthcoming a. "Views of my fellows thinking."

Travis, C. Forthcoming b. "Their work and why they do it."

Further Reading

Travis, C. 2008. *Occasion-Sensitivity: Selected Essays*. Oxford: Oxford University Press.

Travis, C. 2011. *Objectivity and the Parochial*. Oxford: Oxford University Press.

Travis, C. 2013. "The preserve of thinkers." In *Perception: Essays After Frege*, pp. 313–363. Oxford: Oxford University Press. A slightly different (and newer) version is at http://kcl.academia.edu/CharlesTravis (accessed July 22, 2016).

7

On the Linguistic Status of Context Sensitivity

JOHN COLLINS

1 Introduction

The place of context in an overall account of the semantic properties of natural language has been a dominant theme in the philosophy of language over the past 30 years. Indeed, on one way of talking, the topic is coeval with the general study of pragmatics in the various branches of linguistics. Thus it is that philosophy of language and linguistics are increasingly hard to tell apart. My aim in the present chapter is not to survey the field, nor to justify the general concern with context, although I shall indulge in a bit of both; rather, I hope to arrive at some tentative conclusions about the likely *linguistic* status of context-sensitive semantic properties.

There is a broad consensus among theorists from diverse fields and traditions that the contexts in which tokens of linguistic types are uttered systematically contribute to how the tokens are understood in ways speaker/hearers presuppose. That is, it is part of linguistic competence to exploit stable features of the occasion of linguistic acts to communicate successfully. The question I shall address is whether such competence is encoded as invariant 'open' features or variables of syntactic and semantic items or phrases. For example, the place a speaker is situated in when she makes an utterance often bears upon what the speaker says; for that relation to be linguistically encoded is for the linguistic material that delivers such a location-sensitive content to contain an item that ranges over potential locations as an aspect of its standing meaning. To give the conclusion now, I shall answer in the negative, which is just to say that the role of context in contributing to the literal meaning of utterances (*what is said*) is, in general, not itself licensed by lexical or syntactic mechanisms or items, but is, rather, due to extra-linguistic effects, on the assumption that there is no language-specific mechanism of pragmatic interpretation, but only an interpretation that goes by way of general cognition, or at least extra-linguistic cognition. Call such a view *linguistic pragmatism* (Chomsky, 1977; 2000; Lahav, 1989; Recanati, 2004; 2010; Pietroski, 2003; 2005; Travis, 2008; Neale, 2005). My notion of 'a linguistic license' is equivalent to

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Recanati's notion of a 'bottom-up' determination of what is said by way of a fixed interpretation of the linguistic material, which stands in contrast to 'top-down' processes, which go by way of extra-linguistic processes. Thus, pragmatism is the claim that what is said is not bottom-up determined by the language alone, but is a function of both stable linguistic properties (bottom-up) and general, extra-linguistic means of interpretation (top-down).

I do not pretend that I can establish such a broad thesis here, and I shall not dwell on the various ways of spelling out the thesis and the myriad challenges to any such version of the thesis one may entertain. Instead, I shall argue that pragmatism is fully aligned with a standard approach to syntax, and should be the default view of the notion of a linguistic 'context,' *viz*., context is not a well-behaved linguistic notion, but rather a potentially open-ended way of marking the role extra-linguistic factors can play in fixing what is said on an occasion of the use of a linguistic type. In effect, the role of context in determining what is said is a feature of top-down processes. I shall focus on apparent covert context sensitivity in weather predicates as they occur in reports such as *It's raining*.

2 Terms of Debate

2.1 Meaning, Truth-Conditions, What Is Said*, and so on*

Let *meaning* designate whatever semantic properties a linguistic expression (word, phrase, sentence) invariantly possesses in the sense that such properties make a constant contribution to the understanding a competent user of the expression exploits. *Pro tem*, let meaning in this sense be potentially a very thin notion, perhaps no fatter than a set of syntactic constraints on interpretation. A syntactic constraint is a property of sentence that can be specified formally, without reference to general conceptuality, but which allows and disallows certain interpretations. Many examples will follow.[1] The crucial point is that *meaning* is a purely linguistic notion, a property of the language itself that is not determined by context of use or what a speaker intends to communicate; indeed, we might say that meaning is what speakers' intentions cannot trump, but default track. This is why meaning is invariant.

I take utterances, not sentence-types, to have truth-conditions. An argument for this claim is required, for it can seem that *if* meaning is invariant truth-conditional contribution, then that a sentence-type has a meaning suffices for it to have truth-conditions. So-called *minimalists* claim just that: sentence-type meaning suffices for invariant truth-conditions, regardless of whether such a content is apt to be communicated by a token of the type (Borg, 2004; Cappelen and Lepore, 2005). Whatever the truth of this claim, the inference advertised is invalid, that is, minimalism does not get to be true merely by a deft arrangement of distinctions. The contribution a constituent item of a sentence makes to the sentence's truth-conditions may fail to be determinate. In this scenario, linguistic meaning will underdetermine truth-conditions. This claim, that admits great variation, is typically dubbed *contextualism* or pragmatism (I prefer the latter label as already noted) on the understanding that the extra-linguistic ingredient that, along with linguistic meaning, fixes the truth-conditions for an utterance is some relevant selection of contextual factors to which speaker-hearers are sensitive (person, time, addressee, etc.). I shall argue in the following that a fairly radical version of pragmatism is correct, and can be upheld on strictly linguistic grounds. For the moment, I am merely flagging the position as one that is consistent with the existence of literal linguistic meaning *and* truth-conditional content, the position just does not identify the two.

Content (or truth-conditions) is often expressed in terms of *what is said* by an utterance, which stands in contrast with what is communicated, which might involve all kinds of implicature, irony, and other figural effects. *What is said* is what a speaker can intend another linguistically competent hearer to understand without any information other than the shared situation and a shared language. Of course, it doesn't follow that, on any given occasion, communication will be successful, or is even a notion that admits precision where natural language is concerned. Still, it is a truism that knowing the meaning of a sentence *s* (being linguistically competent with respect to *s*) allows one to say *p* with *s* such that one can expect any suitably placed person to understand that one said *p*. This is the notion of what is said, and can be recognized as a crucial aspect of language independent of the epistemology and metaphysics of communication and linguistic publicity.[2]

So, here is a rough picture of the relevant semantic distinctions. Utterances have truth-conditions, which amount to what the speaker *says* on the occasion of utterance. There is a purely linguistic contribution (meaning) to these truth-conditions, which, as things stand, constrains what is said to some degree in the sense of allowing and disallowing literal construals of the linguistic material without fixing any definite construal. I shall hereafter explore a particular route to settling on what the degree of constraint is by way of appeal to syntactic structure. First, let us fix on what is and can be meant by *context*.

2.2 Context

There are (at least) four respects in which the notion of context might be employed in regard to an utterance. First, there is the perfectly general way in which more or less anything may be related to some particular circumstances. I take such a notion of context to be of no peculiar linguistic interest. Context here merely serves to present the utterance in a certain setting without its content being affected, either as a whole or in regard to specific lexical items or constituent phrases. Of course, a specification of a context will trigger various presuppositions and implicatures, but these may be triggered anyway by mutual knowledge without any linguistic aid. This language-independent aspect of context in the most general sense stands in distinction to the following three notions of context.

A second notion of context is familiar from work in formal semantics going back to Montague's (1974) seminal writings and advanced most fully by Kaplan (1989). A context here is essentially a formal abstraction, which we can represent as an *n*-tuple of items drawn from the 'world' such that each item serves as a value for a constituent of the utterance or affects the overall content of the utterance (when suitably regimented). Thus, traditionally, speaker, addressee, time, place, *et al*. find their way into the context and serve as values for (*inter alia*) pronominal items and various functional morphemes (such as tense). A great deal has been written about this idea of a context; my concern, though, is not to renounce or finesse the notion, but largely to take the idea for granted, or at least what I take to be the fundamental insight underlying the idea. If we assume that one way of specifying the content of an utterance is to provide its truth-conditions, then such truth-conditions cannot be specified solely from the knowledge of the meaning of the linguistic type. Much if not everything we say with our utterances is variable in depending on factors that potentially shift from one tokening to another of the linguistic types (tense, personal pronouns, etc.). The present idea of a context, therefore, is simply a way of generalizing over such variability. In this light, a context is simply a component of the truth-conditions of utterances. So understood, it really doesn't matter if we theorize a

context as a set-theoretical object (an *n*-tuple) or let the information be specified as the antecedent of a conditional, with its consequent being the truth-conditional specification of the contextually specified linguistic type. That difference is one of internal machinery in the semantic theory, not a difference in the phenomena. Both approaches have the burden of saying just what information or values enter into the context. The fundamental insight, therefore, is that utterance content is variable over contextual factors in a systematic way under a linguistic license in the sense that values of linguistic items must be in the context, if the utterance is to have truth-conditions.

The items that go into a context in the above sense are those that obtain on the occasion of utterance. There are other variable extra-linguistic factors, though, that bear on truth-conditions but do not obtain on the occasion of utterance. This is a third sense of *context* that pertains to how utterances are assessed or evaluated. For example, a future tense utterance (*Bill will die before he is 80*) will be made true or false by future events, not those that presently obtain; thus, the truth-conditions of such utterances cannot be fully specified in terms of the here-and-now context of utterance alone, but must also include elements of the type that can feature in a context, but have relevant tokens that do not obtain on the occasion of the utterance (e.g., a future time that fixes the truth of a future tense utterance is distinct from the time of utterance).

Appeal to this third notion of context often goes under the label of *relativism* and there is much dispute about whether the context sensitivity of various classes of expression is due to the (here-and-now) context of an utterance or that to which its truth-value assessment is relative. This chapter will not dwell on the standing of so-called relativism either in its own right or as contrasted with some species of 'contextualism' or 'pragmatism.'[3] My only concern is for the potential linguistic licence of the contextual valuation of linguistic material, regardless of whether the relevant contextual factors pertain to the here-and-now or not. I shall, though, turn to one relativist rejoinder to pragmatism in §6.

A fourth sense of *context* is more resistant to a precise specification as it pertains not to specific features of a shared situation or a recognizable point of assessment (a judge or an experiencer), but to a shared understanding of how a word or phrase might be intended. *Polysemy* is an oft-used term for the phenomenon I have in mind, but I here do not intend to prejudge the standing of the effect in terms of its scope and possible constraints. It seems, at any rate, that every predicate save for those with a stipulated meaning is subject to a wide variety of interpretations any of which might be literally meant if the occasion is obliging. Take, for instance, Travis's (2008) much-discussed example of *The leaves are green*. As a photographer, one might be interested in leaves that merely look green, including painted ones (Antonioni had a field in a south London park painted green for *Blow-Up*), but not as a botanist. The point here is not that *green* and other color predicates are lexically ambiguous – they appear not to be – but that, in general, things *count as* satisfying a predicate or not on the basis of shifting understandings of what would count as the thing being so and so on a particular occasion. It is moot whether the variety is open-ended and, if not, what the constraints are on what can count as what.

As advertised, I shall conclude that the fourth sense of context is the one generally appropriate for the determination of what is said. I shall not argue for this directly, though, but only via consideration of the problems that beset the thought that context is a linguistically encoded feature up to the fixation of what is literally said with an utterance.

2.3 Syntax and Context-Sensitive Truth-Conditions

Let us assume that context in at least one of the latter three senses adumbrated above enters into the determination of truth-conditions, and so is a phenomenon an adequate semantic theory for natural language must apparently accommodate; that is, to know a language is to know how to exploit context to say this or that, or, if you prefer, to know that context is exploitable in such and such ways in order to say this or that. My focus will be on the putative syntactic accommodation of contextual factors. To finish off my outline of the terms of the debate, therefore, I shall say something very general about syntax and its relation to truth-conditions.

Naïvely, one might start with the idea that the possessor of truth-conditions is a sentence-type individuated by surface syntax. It is common ground, however, that this will not do for a range of familiar reasons, including what I shall call *overtly context-sensitive expressions*, such as personal pronouns, temporal and spatial adverbials, demonstratives, and so on, which were the chief concern of Montague (1974) and Kaplan (1989). The problem is that what a personal pronoun such as *he* contributes to the truth-conditions of its host sentence will be determined by who the speaker has in mind to refer to by *he*, which will not be invariant over uses of the type. Thus:

(1) 'He is tall' is true iff he is tall

does not record the truth-conditions of the quoted sentence, for it has no determinate truth-conditions in the absence of an identification of who *he* refers to, which the right-hand side of the equivalence does not do. Otherwise put, the form of (1) does not in general fix the theorist's *he*, as it were, as being co-referential with the quoted *he*. There are ways of remedying the problem, as we shall see, but however it is achieved, some departure from the naïve approach is entailed. Let us go the whole hog, then, and say that truth-conditions are a property in the first instance of a *logical form*, leaving it open for which constructions, potentially every one, a valuation of items of the form by context is required for the form to possess truth-conditions, that is, logical form alone may or may not fix truth-conditions by itself.

> *Logical form*: The logical form corresponding to a sentence/utterance is a structure that disambiguates a surface form and licenses determinate values and composition principles for the constituents of the structure such that they make an invariant contribution to the truth-conditions of sentences/utterances of the form.

Traditionally, some such notion of logical form has been borrowed from various systems of formal logic and applied to natural language without real interrogation of how logical form relates to surface form or natural language more generally; that is, logical form is the theorist's language in which the phenomena of meaning as found in the language are described or explained. There is a more constrained conception, however, where 'logical form,' or 'LF,' refers to a level of syntactic structure that realizes the relevant properties of the logical structure under conditions that pertain to natural language syntax generally (May, 1977; 1985; Chomsky, 1977; 1981; Higginbotham, 1985).[4] It is this latter notion I shall have in mind hereafter. Although many of the issues to be broached will not turn on the difference between the traditional notion and the one that prevails in syntactic theory, the distinction is crucial in the following sense. If there is an independent notion of syntax that

enters into fixing possible truth-conditional contributions of linguistic items, then if a notion of logical form cannot be identified with a syntactic level in this sense, then an interface issue arises of how the theorist's logical form is tracking the syntactic properties. Put epistemologically, the best, or at least a very good, justification for a theoretical claim of logical form is precisely that it is evidenced in the syntax. Excuse the abstraction; concrete examples will follow.

Our concern, then, is just this: What is the general relation between logical form structure and context-sensitive truth-conditions? Answering this will give us a fundamental take on context sensitivity as a semantic property of natural language. The short form now is that context sensitivity is not a syntactically coded effect, but lies outside of language proper based upon speaker intention.

3 Overt Context Sensitivity

Let us assume that expressions of a language divide between those that take different values relative to certain contextual parameters and those that take constant values. The sharpness of such a distinction is moot, but there appear to be paradigm cases of each class, such as personal pronouns and spatio-temporal adverbials (*here*, *tomorrow*) on the context-sensitive side and quantifier expressions (*every*, *some*, etc.) and other functional morphemes (tense, *-ing*, etc.) on the other side. Proper names are a controversial case, but let's assume that they are context-insensitive. An adequate semantic theory would feature primitive clauses of some form or other whose content is simply that the value of the name is some object (leaving to one side ontological scruples about what an object is). For example:

(2) a **v**(Bob, *Bob*)
b $[[\text{Bob}]] = \lambda x[x = \text{Bob}]$

It is a common approach to treat pronouns of various stripes as if they were names, save relative to a context. So, whereas *Bob* picks out Bob come what may, a pronoun such as *he* also picks out some definite object, but only relative to a given occasion of its use. The idea is intuitive enough and does reflect features common between names and pronouns, such as rigidity and neutrality over background information. There also appears to be a straightforward way of implementing the idea. In essence, the proposal is that a context-sensitive item is akin to a free variable syntactically encoded via an index and assigned a value relative to an assignment, not as a constant function. Something along such lines is the current orthodoxy in the semantics literature (cf., Montague, 1974; Kaplan, 1989; Fiengo and May, 1994; Larson and Segal, 1995; Heim and Kratzer, 1998; Büring, 2005). Let me first flesh out the view, focusing on a proposal by Larson and Segal (1995), then I shall turn to its problems as a semantic proposal about a fixed linguistic mechanism licensed by logical form.

Following standard post-Tarski (1936/1983) approaches for the semantics of formal languages, variables are taken to be numeral indexes (or to have such indexes – the difference doesn't matter) such that a variable takes as value whatever object is in the ordinal position that is the same as the numeral index of a sequence of objects. So, a variable takes a value relative to sequences of objects drawn from the domain, where variables are distinguished by their indexes, but all remain equally general because, for every sequence position, every object occurs in that position in at least one sequence. The principle works much the same with assignment functions

that dispense with sequences, that is, both methods deliver the same truth-conditions. There is an obvious problem, however, with treating pronouns as free variables. Free variables have a universal or indefinite semantic significance in the sense that, for every object *o*, there is a sequence σ that has *o* in its *i*th position, for all positions *i* (*mutatis mutandis* for assignment functions). A pronoun, however, is not so general: on a given occasion of use, it picks out a definite object. The problem might be put as follows: a variable does not by itself differentiate between objects that might serve to satisfy the host expression – the variable is effectively syncategorematically interpreted in union with its host predicate. Larson and Segal (1995), in response to a problem of this kind, propose that the speaker of an indexical, such as a pronoun, selects a sequence that contains the relevant objects in the relevant positions, but this is a metaphor in the sense that the speaker does not select a sequence, but merely the relevant objects. The same problem arises for assignment functions. Larson and Segal appeal to sequences, I take it, because they capture the generality of indexicals at the type level; the problem is then how to resolve that generality at the token level, a problem that *cannot* arise for variables as satisfied by sequences. Going over to assignment functions does not help. In effect, the model would be one of ambiguity resolution, where the indexical is ambiguous between assignment functions; what the speaker's intention provides is a resolution to the ambiguity in terms of a selection of a unique assignment function from potentially infinitely many options. The problem now, of course, is that a type indexical is precisely not ambiguous as a linguistic item; it is, rather, simply indefinite, which is precisely what sequences capture. The thought that a speaker sheds, as it were, infinitely many assignments in order to select just one assignment faces much the same problem as the sequence approach, for it remains unclear why any options whatsoever are entertained save for the immediate option of the referent being fixed by the speaker's intention. What this misses is the generality of the indexical, but that is missed by selecting a unique assignment too. The generality and definiteness of an indexical can't be squared by treating the indexical as if it were a variable, for a variable is never definite.[5]

There is another aspect to the problem of treating pronouns as variables. We assumed that pronouns come with indexes. Starting with Chomsky (1965), the use of indexes in syntactic theory became the norm, and they certainly are a useful way of encoding deictic and bound readings of pronouns; that is, an index unmatched with another index is read as getting a value from the context (deictic) and an index that is matched with a higher index has a value that covaries with the higher index (bound).[6] Thus:

(3)a Every philosopher loves himself/herself
b Every philosopher thinks he/she is a genius

(3a) is unambiguous with the single Quinese paraphrase:

(4) Every philosopher *x* is such that *x* loves *x*

(3b) is ambiguous between a bound and free construal of the pronoun:

(5)a Every philosopher *x* is such that *x* thinks *x* is a genius
b Every philosopher *x* is such that *x* thinks *y* is a genius

The value of the pronoun on the free construal is contextually determined on the occasion of a tokening of (3b), Kant or Plato, say. The difference may be codified via indexes:

(6)a $[\text{Every philosopher}]_i$ thinks he_i is a genius
b $[\text{Every philosopher}]_i$ thinks he_j is a genius

Here is not the place to venture into the complex field of binding, but it is well to rebut one potential line of reasoning. The dilemma posed for the standard approach of identifying indexical expressions with free variables is that it is the speaker's intention to refer to *x* that fixes *x* as a definite value of a pronoun, which is not a feature that is encoded semantically or syntactically. To be sure, one can stipulate a relation of the kind Larson and Segal propose, but it merely recapitulates the extra-linguistic intention rather than grounds or constrains it. One might think, however, that since pronouns and other complex phrases syntactically feature indexes, then the job of a theory is just to provide values for the indexes in the relevant manner. In other words, the free or bound status of variables *is* encoded in syntax via indexes, so the identification of pronominal context sensitivity with free variability is not sunk, even if the semantic valuation does proceed by way of speakers' intentions. The problem with this thought is the assumption that indexes are kosher syntactic features in the relevant sense.

On the face of it, indexes are the *theorist's way* of marking joint or disjoint construal rather than an independent feature that constrains or fixes what construals are possible. If it were otherwise, then an assignment of indexes to the relevant expressions within a structure would be independent of any intention of joint or disjoint construal. In that circumstance, however, there is no way of fixing one index to be the same as or distinct from another index within a structure just by the properties of the individual indexed items or the structure as a whole, *where* there is an option for joint or disjoint construal. Where there is no such optionality, then indexes are redundant anyway, assuming that some such interdiction is determined by independent structural conditions. Thus, indexes can mark optional construals, but not independently fix one option over another.

The ultimate truth of binding, as it were, is a matter beyond my present scope. The current conclusion is merely that the facts of context sensitivity do not simply fall into line with a naïve theory of indexation that treats pronouns just like free variables. What a speaker intends to refer to is ground level for a *general* account of what pronouns contribute to what is said by an utterance. A theory of meaning in the narrow sense of a theory that specifies what language alone contributes appears to do no more than delimit options of reference and fix necessary construals, all of which appears not to rely upon a system of indexation.

4 Covert Context Sensitivity

Context sensitivity is not restricted to the construal of overt linguistic material. It also appears that the construal of tokens of many constructions depends upon contextual factors that are not explicitly encoded in linguistic material; that is, *if* pronouns are explicit variables, then there appear to be implicit or covert variables as well. *Pro tem*, we may specify the idea in perfectly general terms without appeal to any particular theoretical assumptions. Let us borrow some terminology from Perry (1986), but employ it for our own ends:

> *(UC) Unarticulated constituents*: The stable propositions expressed by sentence tokens (relative to context) of particular types include constituents that are not values of any overt linguistic material occurring in the token.

UC leaves open the provenance of the relevant constituents, whether, that is, they are values of *covert* syntactic variables, semantic effects of lexical meaning, or features of pragmatic

enrichment. UC is useful simply for picking out a range of otherwise diverse phenomena. So, consider weather reports, such as *It's raining*. Tokens of the type appear to express propositions that feature locations in the sense that the truth-values of the tokens depend on how things are at some particular location, that is, some other locations are excluded as truth-conditionally irrelevant. Also consider so-called *predicates of personal taste*, such as *Liquorice is tasty*. Tokens of the type express propositions that concern a particular person's taste, that is, the truth-value of the token turns on how some person judges liquorice as opposed to how liquorice is anyway. Many other examples could be offered, but these cases serve to exhibit some general features that will be my concern.

First, I shall offer some general grounds for why the bare phenomena UC directs us towards should not lead us to populate syntax with covert items that may thus be valued by the unarticulated constituents of the propositions expressed. I shall then consider weather reports in particular as a case that supports this general claim. To finish, I shall briefly consider a relativist rejoinder to the considerations offered.

4.1 The Syntax of Covert Items

A central feature of generative syntactic theory is that it posits covert items and relations in its analyses of linguistic material. According to the approach, how a sentence appears in terms of its phonology/morphology is a set of properties of a more complex structure organized by general principles it is the task of the theory to specify. On the face of it, UC seems to be a prediction of the generative approach, for on the assumption that syntax places structural constraints on what a sentence may mean, then such constraints need not be overtly encoded. Thus, if weather reports are constrained to be locative, and no locative phrase need explicitly occur in a weather predicate for it to occur in a grammatical sentence, then it might well be part of the covert syntax of the general clause type to feature an item that is locatively construed (*mutatis mutandis* for other cases). Even though such reasoning is fine as far it goes, real caution is required if one is minded to generalize the reasoning to every case going, as Stanley (2000) explicitly commends. There are two basic desiderata that need to be satisfied, what I shall call *syntactic license* and *semantic constraint*.

> (SS) A would-be syntactic item must be licensed in the sense that it invariantly affects the grammatical and interpretive status (if distinct) of the structure as a consequence of a general principle.
>
> (SC) A semantic construal of a structure counts as evidence for the structure having a syntactic feature F only if the construal is necessary.

In simple terms, SS enshrines the thought that syntax is not a receptacle for whatever semantic features are deemed stable; a syntactic item must bear upon the *syntactic* status of the host structure. Methodologically, of course, SS operates only in light of some syntactic assumptions, which may always be questioned, but the basic idea is uncontroversial, for if the presence of a covert item did not affect the status of the structure, there would be no syntactic basis for the item at all. Put differently, SS demands a syntactic rationale that might be distinct from any semantic rationale. For example, it is standard to treat infinitival clauses as occurring with a covert subject PRO, which may be bound or 'controlled' by a higher phrase, if embedded, or otherwise be 'arbitrary.' Thus:

(7)a Mary$_i$ tried [PRO$_i$ to leave]
b Mary asked Sam$_i$ [PRO$_i$ to leave]
c PRO to smoke is foolish

Assume a principle that states that a full verbal projection features all of the verb's arguments, even if some or all of them end up occurring overtly outside of the projection. The positing of PRO is one way of satisfying such a principle that has its role to play in explaining a host of phenomena. There is, though, an apparent independent semantic rationale for positing PRO, *viz*., the infinitival has an understood subject. For example, with *Mary tried to leave*, we cannot understand there being no agent of leaving, as if someone's leaving is not even in the offing – that is gibberish. So much, though, does not explain the properties of PRO. One might think, for example, that PRO is like a variable, occurring bound in (7a–b) and free in (7c); such would satisfy the semantic rationale and align PRO with overt pronouns. As it is, though, arbitrary PRO does not occur bound, even when embedded:

(8) Mary asked Sam whether [PRO to smoke is foolish]

(8) does not express Mary's asking whether Sam's smoking would be foolish. The status and interpretation of PRO remain controversial; the current point is simply that positing PRO has a syntactic rationale independent of its construal, and that properties of its construal appear to be peculiar in the sense of not being aligned with overt phenomena.

SC offers a related constraint: even if one is solely considering semantic evidence, the relevant readings must be necessary ones, for the syntax itself is invariant by assumption. For example, (7a) has the single reading under which Mary is doing the leaving, as opposed to an arbitrary other person. Such evidence tells us that the subject position of the infinitival cannot be occupied by something akin to a free variable. As another example, consider familiar adjunct ambiguity:

(9) Mary saw the elephant from Africa

The adjunct prepositional phrase *from Africa* may modify the DP *the elephant* (the elephant is African as opposed to Indian) or the VP *saw the elephant* (Mary may have seen the elephant – African or Indian – across the Israel–Egypt border). The adjunct *cannot* be read as modifying the subject *Mary*. Such a restriction on possible readings tells us that the adjunct is merged in the verbal domain and so cannot selectively modify the subject that occurs outside of the domain (at the 'surface').[7]

There are, to be sure, trickier cases. Consider optional binding phenomena discussed above, such as:

(10) Bill thought he was a genius

(10) is two-ways ambiguous as (9) is. The ambiguity alone does not tell us if the two readings are syntactically encoded, let alone how they are. Remember: SC only says that semantic phenomena count as evidence for syntactic structure, if the phenomena mark a necessity; it does not say that all readings must be syntactically accommodated. So, the ambiguity of (10) might be syntactically unresolved, or be so resolved in ways that do not treat the pronoun as a free variable.

Let us now apply the morals adumbrated to the case of weather reports. Many different phenomena are germane, even with so thin a diet; I shall focus on the most salient data.

4.2 It's Raining *Again*

On the 'standard view,' as Recanati (2010) calls it, the problem of weather reports is how to explain their obligatory definite locative construal; that is, why, for example, *It's raining* is construed as being about a definite location, as opposed to somewhere or other (indefinite) or no particular place at all (*punkt*). The answer provided by the standard view is that the relevant lexical entries contain a locative argument position, a covert variable on some readings, which is valued for locations. So, for example:

(11) $\lambda l \lambda e[\text{rain}(e, l)]$

This would-be lexical entry for *rain* tells us that *rain* picks out an event occurring at a location. Notwithstanding real disagreement over detail, at least Perry (1986), Stanley (2000), Taylor (2001), Corazza (2007), and Neale (2007) agree that weather reports are construed as definite locatives as a matter of lexical understanding, or fixed truth-conditions, not pure pragmatics. In some sense, therefore, they are all committed to (11) as the right lexical entry.[8] In fact, though, no one serious could possibly think (11) records the relevant property of the lexical entry of *rain*, for *rain* is obviously usable outside of weather reporting, where no location at all, definite or indefinite, is relevant. Still, the thought is that a locative argument is saturated in weather reports as a matter of the fixed logical form of the construction rather than pragmatic enrichment (an argument or position is said to be saturated if its value is required for the interpretation of the host predicate). So, ignoring tense, the logical form of *It's raining* is:

(12) $(\exists e)[\text{raining}(e) \wedge \text{Location}(e, l)]$

The kind of analysis exemplified in (12) has at least two things going for it. First, like an overt pronoun, we can take the locative to be contextually valued, which would explain its putative obligatory definiteness: if the argument is contextually valued, then it must be valued as a definite location. Second, again like an overt pronoun, the locative argument can apparently be bound. Consider:

(13) Wherever I go, it rains

This appears to have the obligatory reading

(14) Every location l is such that, if I go to l, it rains in l

Against the standard view, Recanati claims that the lexical entry for *rain* lacks any argument position other than an event position invariant over all predicates:

(15) $\lambda e[\text{rain}(e)]$

A relevant weather report, therefore, might consist simply of the existential binding of the event variable (ignoring tense):

(16) (∃*e*)[raining(*e*)]

The immediate problem with this proposal is that it fails to capture the apparently obligatory definite construal of weather reports. When one utters *It's raining*, one surely doesn't *say* that there is a raining event taking place (somewhere or other). One cannot help, it would seem, but to speak of one's present location; or at any rate, a location otherwise contextually salient, such as the location of one's addressee. For precisely this reason, as explained above, the 'standard view' represents the lexical entries of meteorological predicates as containing a locative variable. Recanati offers a complex response to this challenge, but here I shall only be concerned with his claim that scenarios are readily imaginable where weather reports have indefinite and even *punkt* construals, just as (15) predicts.

Recall that, according to the 'standard view,' meteorological predicates have a locative argument position that must be filled, if an utterance hosting the predicate is to be semantically acceptable. Such a condition would explain the definite construal of weather reports being apparently mandatory. The view, though, appears to make a strong prediction: *punkt* construals of weather reports are unavailable; and while indefinite construals might be available, they would have to involve a quantification into the locative position. Recanati (2002; 2004; 2010) offers the following kind of scenario as a counter-example to this prediction:

The Weatherman Scenario
The Earth has suffered a massive ecological catastrophe, the chief consequence of which is that rain no longer falls. The remnants of humanity have decamped to an underground bunker. Fortunately, before the survivors were forced from the surface, they placed sensors all over the planet in order to detect the hoped-for future rainfall. The detection mechanism, though, is not that sophisticated; the console in the bunker's monitoring room, manned by the weatherman, lights up just if rain falls on any of the sensors, but it doesn't immediately record which sensor is so affected. To figure out the identity of the relevant sensor requires lots of laborious calculation. One propitious morning, the light on the console begins to flash. The weatherman excitedly cries to his colleagues, 'It's raining!'

The intended intuition elicited here is that the content of the weatherman's utterance is *punkt*, for neither he nor his audience know of the location of the rain; for sure, they know that it is raining *somewhere* or other, but this reveals some knowledge about the nature of rain, not what is saturated in the linguistic content. The weatherman, we may say, has evidence of rainfall, but is not yet making a claim about its location. Thus, the logical form of the weatherman's report would appear to correspond to (16), which lacks the putative obligatory locative argument position.

It is important to note that the weatherman scenario is *not* intended as an argument for the possibility of it's raining at no place at all, as if there could be locationless rain, as it were. If it is raining, then it must, indeed, be raining somewhere, whether or not one knows where. The point of the scenario, at least in my hands, and Recanati's (2010, pp. 88–90) too, I take it, is to show that there are *punkt* readings of weather reports that appear to require no variable binding. Thus, in the absence of a good reason to think that some kind of covert quantification is in play, the scenario militates for (16) being the appropriate logical form. The fact that the weatherman would acknowledge that it must be raining *somewhere* does

not, in this light, belie (15) as a depiction of the lexical content of *rain*. It is the weatherman's wider knowledge about rain that licenses the inference to its raining somewhere, not an aspect of the lexical content. His understanding of the metaphysics of rain – that it is locational – overrides the mere semantic content that there is a raining event as determined by the fixed semantic properties of the linguistic type. That it is raining iff it is raining somewhere does not entail that *It's raining* features a locative argument as an aspect of its syntax or semantic content. There is much else to say here, but it is best delivered by way of the consideration of an objection.

As presented so far, Recanati's weatherman reasoning suffers from a decisive objection. The weatherman scenario is perfectly coherent, but the weatherman's weather report is not genuinely indefinite, let alone *punkt*. To be *really* indefinite, the truth-conditions of the utterance must be sensitive to every location: if it is raining *anywhere*, then the utterance is true. If some set of locations is constitutively excluded as being potentially determinate of the truth-value of the utterance, then the complement of the set of such locations will constitutively include all and only the truth-determining locations. So, in a broader sense, the utterance will not be indefinite, still less *punkt*, but definite relative to the relevant complement set of locations. In concrete terms, the weatherman's utterance is definite as regards locations on Earth, for neither the weatherman nor his audience would take the weather on another planet to be truth-conditionally relevant. The weatherman's utterance is definite after all, albeit in a broader sense than we might at first imagine. Such considerations suffice, I think, to confound the weatherman scenario's import as presented, but they invite a recasting of the scenario.

One can recast the scenario in two different ways that undermine the objection just raised. First off, consider a scenario just like the original weatherman one, except that the weatherman is entirely clueless about where he is. As far as he knows, he could be on Earth, Mars, or in a spaceship moving at near light speed; just so, he hasn't a clue where his sensors are scattered. In this scenario, the weatherman may still legitimately cry, 'It's raining,' when his console flashes, but there is just no definite location, no matter how expansive, that may serve as the location that satisfies the weatherman's report; at any rate, the weatherman certainly cannot exclude any location as truth-conditionally irrelevant for his utterance, for he doesn't know his own location or the location of the sensors in order to exclude any other location. This 'ignorance' scenario, I think, clearly shows that weather reports do not need to be construed definitely.

Second, consider an 'intergalactic scenario,' just like the first weatherman scenario save that the sensors are dispersed throughout the universe, not just on the surface of the Earth. If the objection to the initial weatherman scenario is merely that the weather report is definite after all, because the Earth constitutes the definite value of the weather report on the grounds that non-terrestrial locations are excluded as truth-conditionally irrelevant, then the intergalactic scenario would appear to spike the complaint, for no locations are excluded as truth-conditionally irrelevant. Arguments can be had here. Extrapolating from remarks by Neale (2007, p. 370, n. 77), Perry (2007, p. 545), and Korta and Perry (2011, pp. 112–113), one may suggest, for instance, that the entire universe remains a location, so the intergalactic scenario still doesn't provide a case where the weather report is construed as lacking a locative argument position.[9] The force of this objection is questionable.

The intergalactic scenario satisfies the just demand that a non-definite construal shouldn't exclude any space–time point as being truth-conditionally irrelevant. The riposte now is, in effect, that every location together amounts to a location. This is not a silly

thought, but nor is it obviously true. Consider two cases. One may imagine a physicist saying, 'It is cold here,' intending to speak about the temperature of the universe. Of course, we also have the age-old philosophical query, 'Why are we here?' One would miss the import of the philosophical query by responding with, 'What, you mean as opposed to there?' The query is not intended to be about some location as opposed to another, but all locations as such. Similarly, the physicist is not refuted by one's appealing to the core of the Sun, for he has in mind not this or that location, but all locations taken as a sum. Even granting, though, that the whole of space–time may constitute a location, the point of the initial complaint is answered: the weather report is non-definite to the degree to which no location is excluded as not being *here*. It would cease to be non-definite were we to imagine some location outside of the universe whose climate was irrelevant to the truth of the weather report. Fine. Now we just change the scenario again to include the multiverse, if you will – and so on and on. In short, establishing a genuinely non-definite reading of a weather report does not involve severing the very idea of a location from the interpretation of meteorological predicates; it suffices that no location is excluded by the reading.

The question remains, though, whether either the 'ignorance' or the 'intergalactic' scenario admits *punkt* construals of weather reports. One line of reasoning for a negative answer is that in both cases we can render the weathermen's respective utterances as *It's raining somewhere*, which apparently amounts to an existential quantification into the locative position of the predicate; indeed, the original weatherman scenario may be so rendered too. So, whereas in the above scenarios the weathermen do not know where it is raining and so do not make claims about that definite location, they still may be understood as making a claim about some or other location, namely the one where it is raining. A defender of the 'standard view' may now claim that the locative position in the meteorological predicates can come in two forms: either contextually valued, as in the quotidian cases, or implicitly existentially bound, as in the *outré* cases just discussed. So, the logical form of the non-definite cases just discussed is rather indefinite, with a bound locative position, instead of being genuinely *punkt*, with no locative position at all.

Independently of further considerations, I readily concede that there is no knock-down argument against this reasoning in favor of the monadic account of *rain* offered by Recanati. That said, the argument suggested bolsters the 'standard view' more than it defeats the monadic view, for the argument as presented simply assumes that the fact that raining must take place at some location is built into the semantics or lexical entry for *rain*. The existence of such a constitutive relation, though, is precisely what is at issue. The opponent of the 'standard view' is not *obliged* to produce a *punkt* scenario, but only to cast doubt on whether readings of weather reports are best explained in terms of saturation of locative positions. Indeed, given the metaphysical connection between rain and its location, the opponent might well happily acknowledge that no scenario could possibly force a clearly *punkt* construal, but by itself that does not show anything about the status of the would-be locative argument position. What is required for progress here is a syntactic or lexical account of the putative locative position. That is, if it could be shown that *rain* carries a locative argument as part of its lexical content, then that would show that the locative metaphysical character of rain is linguistically encoded as opposed to being an aspect of our general metaphysical understanding of raining events that always overrides a *punkt* semantic construal, which is all the linguistic properties fix.

It is also worth noting that the utterances of neither of the weathermen may be properly construed as *It's raining where the signal came from*. The respective weathermen believe that or have evidence for it, but do not *say* that. What they say would be true, and be taken to be

so, if the signal was a false alarm, but, serendipitously, it was in fact raining elsewhere. Remember, the weathermen have no way of knowing just where the signal is coming from; they merely have evidence that it is raining.

4.3 Quantifying In

The second strand to the 'standard view' is that meteorological predicates *must* contain a locative position because we can quantify into it. If there were no such position, then the quantifications would be, contrary to fact, illicit. The relevant data are exemplified in (17) (here, (13) is repeated as (17a)):

(17)a Wherever I go, it rains
b Every location *l* is such that, if I go to *l*, it rains at *l*

The exhibited reading of (17a) is enabled by an implicit position licensed by *rain* being bound by the prefixed apparent quantifier phrase, or so claims the 'standard view.' That is to say, *It rains* contains a locative variable in non-quantificational constructions; hence it is that it can be bound in (17a). If this is so, then we supposedly have a criterion for the presence of a variable item, *viz.*, a bound reading is available. I have depicted the bound reading in (17b) as if the variable associated with *rain* is *simple*. This need not be the case, of course. Instead, following Stanley (2000), one might think of the variable *l* as a stand-in for a complex – $f(x)$ – that is contextually valued in a twofold way.[10] The variable x may be bound or valued contextually, in the present case as a location (an element of the context), and the function f is valued (in context) as a function from the location value of x to the locations where it rains (*mutatis mutandis* for the variable complex putatively associated with DPs). So, $rains_{f(x)}$ takes as its value the set of locations where it rains.

Before assessing the argument offered, it is well to note that the principal cases to be discussed, exemplified by (17a), appear to involve adjunct free relatives (e.g., *wherever I go*), whose syntactic status is the subject of much dispute. We can argue about whether the free relative as it occurs in (17a) binds some kind of argument position, but such a proposal is not underwritten by standard movement conditions, such as quantifier raising (QR) or the like, that are typically taken to be diagnostic of quantification in syntax. This is obvious enough, for the movement is perfectly overt (the relative clause is moved from a position modifying *rain*) and no 'surface'-level properties appear affected by the free relative being fronted or not, that is, it makes no difference to the interpretation whether the relative clause appears at the front of the sentence or as appended to *rains*. In effect, therefore, the syntactic issue is whether the free relative satisfies the selection requirements of the matrix verb *rains*, which if so would provide evidence for a locative argument position, that is, the relative clause would not be an adjunct. That, however, is precisely the issue in dispute, and we have seen no evidence to think that the locative construal of *rain* is a matter of its syntactic projection; that one can append a locative relative clause tells us nothing, since the clause itself may be an adjunct, as it appears to be. It is still open, of course, for one to think that, when fronted, the phrase does bind a lower variable position, and that this is a syntactic affair, but I know of no reason to postulate such an apparently *ad hoc* syntax.

(17) does appear to pattern *rain* with other relational items (nouns, adjectives, and verbs), which lends weight to the thought that *rain* itself expresses a locative meaning however syntactically realized. Consider the following cases:

(18)a For most Arabs, America is the enemy [of the Arabs]
b Everyone prefers a local [to them] bar
c Whenever the copy-editor made a mistake, the proofreader would notice [the mistake]

Questions arise here concerning implicit arguments and their potential syntactic projection, whether in nominal, verbal, or adjectival domains. Note, however, that the bare idea of an implicit argument does not involve its syntactic projection.[11] It is clear, though, and here Stanley is right, that we normally use DPs to quantify over less of the domain than the lexical N covers, that is, the domain is restricted. A natural way of codifying this phenomenon is to think of a DP as being associated with a variable that is contextually valued in a way that restricts the lexical domain. So, when one says *Every boy is in the park* in some context *C*, one means and is taken to mean that every boy, in the relevant sense, such as every boy on the school trip you have organized, is in the park. Talking in this way, however, leaves the status of the variable involvement open in the way acknowledged by Partee (1989), that is, whether the variable is a semantic, pragmatic, or syntactic representation. Stanley (2000, 2007), Stanley and Szabó (2000), Martí (2006), and Schaffer (2011), however, contend that quantificational dependence readings can be evidential of the presence of a syntactic free variable; in particular, the claim that general domain restriction on quantificational DPs is realized by a valued covert free variable is supported by the existence of readings where the putative variable is bound.

I shall here consider an argument against this kind of binding consideration due to Recanati (2004), who doesn't so much reject the consideration outright as doubt that it is as compelling as its proponents imagine. I agree. Recanati suggests that the principle in operation here overgenerates, that is, were a quantificational reading reason enough to posit a covert variable, we should find ourselves positing covert variables where there aren't any. In other words, Stanley's (2000) claim that where we have semantic binding, we have syntactic binding, is far too strong a condition, for it obliges us to posit syntactically realized variables for any verb one cares to mention, any verb that can enter into the kind of context exemplified in (17). There are other fundamental problems, too, but space is limited (see Collins, 2007; Neale, 2007; Elbourne, 2008; Pupa and Troseth, 2011). Still, the overgeneration worry is worth discussion, for it has, to my mind, been rejected for inadequate reasons.

So, consider (19):

(19) Whenever Bill cooks mushrooms, Sam eats

Assume we naturally understand this sentence to mean that Sam eats mushrooms on all those occasions when Bill cooks mushrooms. It would now appear, though, that we have to posit a variable as the object of *intransitive eat* in order for it to be bound – an unwelcome result. The problem is that if we are using the quantificational test to determine whether or not a predicate projects a variable, then we appear to be obliged to posit variables where there is no contextual relativity.

Martí (2006), endorsed by Stanley (2007, pp. 226, 244), rejects Recanati's reasoning on the basis that (19) is not essentially quantificational. Consider the following discourse:

(20)A: Whenever Bill cooks mushrooms, Sam eats
B: #No he doesn't; curiously, he eats something else

The supposed intuition marked in (20) is that B's response is anomalous. In distinction, the corresponding discourse about rain is OK:

(21) A: Whenever Bill lights a cigarette, it rains
B: No it doesn't; curiously, it rains somewhere else

If, the reasoning goes, (19) is a case of quantification into the putative object position of *eat*, (20) should be fine, for A's utterance would, indeed, be false, as B reports, should Sam eat something other than mushrooms. That B's response is (supposedly) anomalous signals that the second clause of B's statement does not refute what A said, which shows that the complement of *eat* in A's utterance is not dependent on the content of the fronted adverbial. In (21), there is no corresponding problem, which is meant to indicate that A's utterance in (21), but not in (20), is genuinely quantificational.

I find this argument unconvincing, to say nothing of the fact that were the difference to obtain, it would remain unexplained, since the syntax of the free relatives would remain invariant, presumably, between the two kinds of case.[12] Recanati's point, or at least the point I wish to make, is not that (19) *must* take a bound reading, only that it may do so. Such a possibility suffices to support the overgeneration criticism, for the claim Recanati is challenging is that bound readings mandate as a matter of saturation a variable position into which a prefixed phrase may quantify. A single case is enough to refute the generalization. That general point aside, B's response in (20) is not necessarily deviant. The response, 'No he doesn't,' is elliptical for 'No he doesn't eat' or 'No he doesn't eat mushrooms,' where these completions correspond to what B may intend by his response. On the first construal, B's response is incoherent, for B is simply contradicting herself, saying that A both eats and doesn't eat. On the second construal, B's response is perfectly fine, but is only enabled if *eat* in A's utterance is construed transitively with its object supplied by the prefixed phrase. The situation is the same for (21). The phrase, 'No it doesn't,' is elliptical for either 'No it doesn't rain' or 'No it doesn't rain wherever Bill lights a cigarette.' Again, on the first construal, B's response is incoherent without further ado, for B is being self-contradictory. On the second construal, the response is fine. The cases are therefore symmetrical; the potential for asymmetry arises from differential construal of the ellipsis. Of course, Martí and Stanley assume that the ellipses are filled in differently in the two cases, which generates the differences suggested, but there is no argument for this that I can discern, for both ellipses admit the two fillings I have suggested.[13] Moreover, symmetry is exactly what we would predict on the basis of taking both kinds of construction to feature adjunct free relatives, as they appear to do, which can be the subject of ellipsis or not along with the verb they modify (in distinction to arguments proper, which must be elided). The remaining question is which filling of the ellipsis is favored, if any, and why, but if what I have said so far is correct, then the answers to such questions will not be determined by syntax, which leaves the options open.[14]

All of these considerations concerning weather predicates have been animated by SS and SC constraints: in effect, unless independent syntactic evidence can be found or the relevant readings are necessary (non-optional), a syntactic explanation of the readings available is precluded. No particular theoretical framework need be deployed to arrive at this conclusion.

5 Whither Context?

If all of the above is on the right lines, how should the notion of *context* be understood? The default view, rather than a nihilistic resignation to linguistic chaos, should be that context, as a determining factor in *what is said*, is not a linguistically kosher notion. It pertains to our general ability to attribute content, which is under linguistic constraints, but essentially undetermined by them. Such a view is argued for on its own terms in different ways by linguistic pragmatists of various stripes. One might be tempted by *relativism* of the kind advertised in §2. Tokens of a sentence-type such as *Liquorice is tasty* appear to be true or false relative to an experiencer or a judge. The linguistic type itself, however, is perfectly well formed without an item referring to such an agent, that is, a speaker needn't append *for me*, say, in order to say that liquorice is tasty for them. According to the relativist, for the relevant class of constructions, truth-values can only be assigned relative to the provision of an 'index' or 'circumstance' that is, itself, not a factor of the content of the sentence-type, but includes 'items' to which truth-value is sensitive, such as an experiencer/judge. In its negative aspect, therefore, relativism is akin to the brand of pragmatism I espouse: *what is said* as a truth-evaluable content is not linguistically licensed. Relativism as a positive doctrine, though, would domesticate context, at least in many of its aspects, in a way that is implausible, at least on linguistic grounds.

Correctly, the relativist rejects the putative linguistic license for the indexation of truth-values to (*inter alia*) experiencers or judgers, that is, the narrow linguistic content need not feature variables ranging over relevant agents. It remains unclear to me, though, what the doctrine amounts to, if understood as a thesis about language. It might be, as Lewis (1980) suggested, that the 'shifty' items of the index should correspond to potential linguistic operators. In the spirit of Lewis, if not to the letter of his proposal, Predelli (2005) and MacFarlane (2009) have generalized the basic idea of a 'shifty' index. Assume that there is a *counts as*-operator, much as there are modal (e.g., *possibly*), temporal (e.g., *yesterday*), and precision (e.g., *roughly*/*strictly speaking*) operators. The would-be operator basically resolves how a predication is to be construed relative to the categorization the speaker intends. For example, *x* can *count as* a red pen in one sense but not in another depending on the circumstance. Truth-value of a sentence in context is assigned relative to the value a coordinate of the index takes under the scope of the operator. There is, however, no general covert *counts as*-operator, or at least not one outside of the intentions of a speaker. So, I do not see relativism as an independent position, *if* construed as a general claim concerning the determination of *what is said*. Relativism amounts to another way of spelling out the pragmatist thought that *what is said* is not licensed by linguistic material alone, but by the intentions of the speaker under linguistic constraints.

A related worry pertains to the covertness of the would-be *counts as*-operator, which I take to be necessary, if relativism is to be rolled out for linguistic phenomena generally. It is clear when modal, temporal, and other adverbial operators apply, at least it is clear how we are to construe the relevant word; this is because they are overt, or have overt markers, such as, perhaps, conditional clauses. In these cases, the 'shifty' aspect is directly licensed by the linguistic material. When the putative operator is covert, the construal of the overt material depends upon a myriad of factors concerning the particular speaker and the topic of the claim. On most occasions, I am simply at a loss with moral and aesthetic claims, say, whether the speaker means to be a relativist or not, precisely because the difference is not linguistically marked. In my experience, it really all depends how seriously the person is interested in

the subject at hand. Even with *tasty* and *fun*, it is not obvious whether speakers are making claims about themselves, in a sense, or properties to which others should be sensitive, too. This is *not* a linguistic matter. If we are to imagine a *counts as*-operator as a free-floating device that applies or not depending on the peculiar concerns or intentions of a speaker directed towards some particular subject-matter, then the putative operator is just a notion that labels the phenomena. I should say that I do not mean to doubt that one can build a relativist semantic theory, or that one may construe the *counts as* notion as a meta-theoretical device that fixes a valuation for a non-relativist semantic theory. My concern is simply that such efforts amount to a theorization about *what is said* that renders it as not wholly determined by the linguistic material alone, which amounts to a form of pragmatism.

6 Concluding Remarks

The interface between syntax, semantics, and pragmatics is such an intriguing area because while we know, or think we know, how the broad distinctions should be mapped onto the actual linguistic phenomena, each case, when looked at in detail, offers a great complexity, which should give us pause over the confidence with which we wield the broad distinctions. My intention in the foregoing has been to settle on what I think is the most stable point of the triad, *viz*., syntax, and see context sensitivity is licensed. From that perspective, albeit in the narrow way pursued here, the conclusion seems to be that context sensitivity is not a feature that syntax encodes; rather, syntax simply fails to fix a proposition, leaving the determination of what is said, especially as that notion is sensitive to prevailing factors of the circumstance of utterance, to extra-linguistic factors to determine. This amounts to a version of pragmatism, which follows, not from a direct consideration of the variety of what can be said with a single linguistic type, context willing, but from a modest view of what language alone properly provides to the fixing of content.[15]

Notes

1 As a taster, just consider the open-endedness in the nominal domain of how to interpret genitives and compound nominals. The syntax will determine that *Mary's car*, say, picks out a car that is somehow related to Mary; the determiner phrase (DP) cannot be read as picking out Mary in relation to a car (the Mary who is next to a certain car, say).

2 The notion of *what is said* due to Grice (1989, p. 87) obeys a syntactic constraint, i.e., what is said is expressed by 'the elements of [the sentence], their order, and their syntactic character.' On this view, *what is said* is a thin notion that tracks syntactic-lexical properties. I use the term more ecumenically: *what is said* amounts to the literal truth-evaluable content of an utterance, which, relative to the sentence-type employed by a speaker, may or may not respect Grice's constraint from occasion to occasion. That is to say, Grice's constraint might not issue in a content *p* that is sayable (cp. Bach, 1994). My arguments do not turn on the subtle matter of distinguishing *what is said* from various related notions. It is a term of art to be defined as is useful.

3 See, for example, Richard (2008) and MacFarlane (2014), among many others, for elaboration and defense of relativist positions.

4 The principle initial motivation for a syntactic level of LF was the explanation of scope taking as a semantic property via movement or displacement of the quantifier DP as an instance of a general syntactic rule that admits the movement of any item (under certain conditions). The correctness of this claim and associated hypotheses is irrelevant to our present concerns; it suffices for present

purposes if we minimally assume an independent syntax that constrains possible truth-conditional interpretation.

5 The argument here is set out more fully in Collins (2014).

6 Here I elide great complexity about the significance of the relations in which indexed items must stand if they are to have certain interpretations, but not others.

7 The condition here is not trivial. Consider so-called *depictives*:

(i)a The doctor examined the patient naked
b The policeman arrested the robber drunk

The cases are ambiguous between who is naked/drunk, which appears to contradict the claim in the text concerning adjuncts insofar as the low adjective in these cases can modify the subject. Note, though, that it is the VP being modified. For example, (ia) does not have a reading whereby a doctor, who is naked, examined a patient at some previous time, where both were fully clothed, that is, the doctor, if naked, must have been naked while examining the patient.

8 Perry (1986; 2007, p. 548) is neutral about how the locative position of *rain* is realized, as are Cappelen and Hawthorne (2007) and Hawthorne and Manley (2012, pp. 117–122). Taylor (2001) thinks the position is lexically encoded, but syntactically inert. Korta and Perry (2011, p. 110) endorse Taylor's view, but consider such lexicalization as 'a social phenomenon' contingent on the peculiarities of given speakers rather than a shared lexical competence. Neale commends Taylor's view on a more cognitive construal, without quite endorsing it. Stanley (2000 and 2007) thinks the locative position is marked by a variable in a projected syntactic position (Corazza, 2007, holds a variant of this view), although he remains neutral on precisely how the position is realized syntactically (cf. Stanley, 2007, pp. 248–249). See Collins (2007) for discussion. If one objects to talk of 'standard views,' then it will not affect my arguments if 'Perry *et al.*' substitutes for the offending phrase.

9 Korta and Perry (2011) and Perry (2007) raise the issue not in response to Recanati's weatherman scenario, but Cappelen and Lepore's (2005 and 2007) 'minimalism,' according to which unarticulated constituents are a 'myth.' So, minimalism has it that *It's raining* always expresses some minimal proposition, which is apparently true just if there is rain anywhere at all, even on Venus, say. To accept the intended import of the intergalactic weatherman scenario, however, is not to endorse minimal propositions. All I intend the scenario to show is that location is not saturational because it is optional. It doesn't follow that every grammatical sentence is apt to express a (minimal) proposition, or any proposition at all. Still, the minimalist and the pragmatist are free to exploit the same scenarios for different ends.

10 Stanley is not committed to his putative variables occupying argument positions within the syntactic projection; instead, he appears to think of them as a species of adjunct or even as 'relative clauses' (some kind of free relative, presumably) (Stanley, 2007, pp. 248–249). This makes his position similar to that of Martí (2006) insofar as the variables will be syntactically optional (for discussion, see Collins, 2007). The precise syntactic position of the putative variable will not affect the following arguments.

11 The syntactic status of implicit arguments remains controversial (see Landau, 2013, for recent survey and discussion). What is clear, at any rate, is that implicit arguments in the nominal domain (arguments of so-called relational nouns) appear to be strictly optional, and so not syntactically projected (cf. Adger, 2013).

12 Bourmayan and Recanati (2013) offer related but independent considerations. Although I am sympathetic to their position, our views diverge on details. Bourmayan and Recanati take the basic lexically encoded meaning of intransitive *eat* to be akin to *eat something*, with the object position existentially quantified over. This is problematic, however, given the thematic constraints on complement deletion, which appear to demand a richer construal than *something* for the deletion case. Moreover, by way of situation variables, they take the unbound reading to be 'literal,' and to 'extend' the situation specified in the free relative, with the bound reading being

produced by 'pragmatic modulation' (Bourmayan and Recanati, 2013, p. 134). I share the intuition of the relative availability of the readings, but am happy to think that neither reading is 'basic.' It is plausible that the free adjunct free relative *whenever φ* has a fixed concessive reading, which leaves open whether the construal is bound or not.

13 Of course, the default syntactic position here should be that, in both cases, the ellipsis is filled by copying the bare predicate (*eat* or *rain*), but that, again, would make the cases symmetrical. So, whatever is explaining the asymmetry appears not to be a straightforward syntactic condition on ellipsis.

14 It is plausible, as a matter of pragmatics, that since *eat* does take an object, if it is used intransitively, it is construed indefinitely, which supports the favoring of an intransitive construal of the ellipsis, whereas since *rain* takes no object, the construal of the location is more open to interpretation. None of this, of course, militates against Recanati's overgeneration complaint.

Stanley (2007, pp. 226–227) offers other examples in support of Martí's 'test,' but they are treatable in a similar fashion, I think. A further consideration is offered by Cappelen and Hawthorne (2007), who correctly point out that Martí's discourse test patterns *dance* and other verbs with *rain*: A: Everywhere Jane went, she danced. B: No she didn't, she only danced in a few places. So, on the assumption that *dance* isn't inherently locative, the binding argument overgenerates, treating *rain* and *dance* as both possessing a bindable covert locative variable. If *dance* is *supposed* to pattern with *eat*, then B's response should be anomalous, as if it were contradictory. My take on this case, as with the others, is that the elliptical clause of B's response is ambiguous.

15 My greatest debt is to François Recanati, for the inspiration of his work and many conversations about the issues addressed above, especially during my stay in Paris as a visiting EHESS professor in 2013. I also thank Adam Sennett, Dan Zeman, Emma Borg, Anouch Bourmayan, Kent Bach, Michael Glanzberg, Paul Pietroski, Bob Hale, and especially Alex Miller, for suggestions about both content and style, from which the chapter has benefited greatly.

References

Adger, D. 2013. *A Syntax of Substance*. Cambridge, MA: MIT Press.

Bach, K. 1994. "Conversational impliciture." *Mind & Language*, 9(2): 124–162.

Borg, E. 2004. *Semantic Minimalism*. Oxford: Clarendon Press.

Bourmayan, A., and F. Recanati. 2013. "Transitive meanings for intransitive verbs." In *Brevity*, edited by L. Goldstein, pp. 122–142. Oxford: Oxford University Press.

Büring, D. 2005. *Binding Theory*. Cambridge: Cambridge University Press.

Cappelen, H., and J. Hawthorne. 2007. "Locations and binding." *Analysis*, 67(294): 95–105.

Cappelen, H., and E. Lepore. 2005. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

Cappelen, H., and E. Lepore. 2007. "The myth of unarticulated constituents." In *Situating Semantics: Essays on the Philosophy of John Perry*, edited by M. O'Rourke and C. Washington, pp. 199–214. Cambridge, MA: MIT Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1977. *Essays on Form and Interpretation*. Amsterdam: North-Holland.

Chomsky, N. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht, Netherlands: Foris.

Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Collins, J. 2007. "Syntax, more or less." *Mind*, 116(464): 805–850.

Collins, J. 2014. "The nature of linguistic variables." *Oxford Handbooks Online*. DOI:10.1093/oxfordhb/9780199935314.013.004.

Corazza, E. 2007. "Thinking the unthinkable: excursion into Z-land." In *Situating Semantics: Essays on the Philosophy of John Perry*, edited by M. O'Rourke and C. Washington, pp. 427–450. Cambridge, MA: MIT Press.

Elbourne, P. 2008. "The argument from binding." In *Philosophical Perspectives 22: Philosophy of Language*, edited by J. Hawthorne, pp. 89–110. Oxford: Wiley-Blackwell.
Fiengo, R., and R. May. 1994. *Indices and Identity*. Cambridge, MA: MIT Press.
Grice, P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Hawthorne, J., and D. Manley. 2012. *The Reference Book*. Oxford: Oxford University Press.
Higginbotham, J. 1985. "On semantics." *Linguistic Inquiry*, 16(4): 547–593.
Heim, I., and A. Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.
Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Oxford University Press.
Korta, K., and J. Perry. 2011. *Critical Pragmatics: An Inquiry into Reference and Communication*. Cambridge: Cambridge University Press.
Lahav, R. 1989. "Against compositionality: the case of adjectives." *Philosophical Studies*, 57(3): 261–279.
Landau, I. 2013. *Control in Generative Grammar: A Research Companion*. Cambridge: Cambridge University Press.
Larson, R., and G. Segal. 1995. *Knowledge of Meaning: An Introduction to Semantic Theory*. Cambridge, MA: MIT Press.
Lewis, D. 1980. "Index, context and content." In *Philosophy and Grammar*, edited by S. Kanger and S. Öhman, pp. 79–100. Dordrecht, Netherlands: Reidel.
MacFarlane, J. 2009. "Nonindexical contextualism." *Synthese*, 166(2): 231–250.
MacFarlane, J. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.
Martí, L. 2006. "Unarticulated constituents revisited." *Linguistics and Philosophy*, 29(2): 135–166.
May, R. 1977. "The Grammar of Quantification." PhD diss., Massachusetts Institute of Technology.
May, R. 1985. *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.
Montague, R. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, CT: Yale University Press.
Neale, S. 2005. "Pragmatism and binding." In *Semantics versus Pragmatics*, edited by Z. G. Szabó, 165–285. Oxford: Oxford University Press.
Neale, S. 2007: "On location." In *Situating Semantics: Essays on the Philosophy of John Perry*, edited by M. O'Rourke and C. Washington, pp. 251–394. Cambridge, MA: MIT Press.
Partee, B. H. 1989. "Binding implicit arguments in quantified contexts." *Chicago Linguistic Society*, 25: 342–365.
Perry, J. 1986. "Thought without representation." *Proceedings of the Aristotelian Society*, suppl. vol. 60: 263–283.
Perry, J. 2007. "Situating semantics: a response." In *Situating Semantics: Essays on the Philosophy of John Perry*, edited by M. O'Rourke and C. Washington, pp. 507–576. Cambridge, MA: MIT Press.
Pietroski, P. 2003. "The character of natural language semantics." In *Epistemology of Language*, edited by A. Barber, pp. 217–256. Oxford: Oxford University Press.
Pietroski, P. 2005. "Meaning before truth." In *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, edited by G. Preyer and G. Peter, pp. 255–302. Oxford: Oxford University Press.
Predelli, S. 2005. *Contexts: Meaning, Truth, and the Use of Language*. Oxford: Oxford University Press.
Pupa, F., and E. Troseth. 2011. "Syntax and interpretation." *Mind & Language*, 26(2): 185–209.
Recanati, F. 2002. "Unarticulated constituents." *Linguistics and Philosophy*, 25(3): 299–345.
Recanati, F. 2004. *Literal Meaning*. Cambridge: Cambridge University Press.
Recanati, F. 2010. *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.
Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.
Schaffer, J. 2011. "Perspective in taste predicates and epistemic modals." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 179–226. Oxford: Oxford University Press.
Stanley, J. 2000. "Context and logical form." *Linguistics and Philosophy*, 23(4): 391–434.
Stanley, J. 2007. *Language in Context: Selected Essays*. Oxford: Oxford University Press.

Stanley, J. and Z. G. Szabó. 2000. "On quantifier domain restriction." *Mind & Language*, 15(2–3): 219–261.

Tarski, A. 1936/1983. "The concept of truth in formalized languages." In *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, 2nd edn, edited by J. Corcoran, translated by J. H. Woodger, pp. 152–277. Indianapolis: Hackett.

Taylor, K. 2001. "Sex, breakfast, and descriptus interruptus." *Synthese*, 128(1): 45–61.

Travis, C. 2008. *Occasion-Sensitivity: Selected Essays*. Oxford: Oxford University Press.

Further Reading

Asher, N. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge: Cambridge University Press.

Partee, B. H. 1984. "Compositionality." In *Varieties of Formal Semantics*, edited by F. Landman and F. Veltman, pp. 281–311. Dordrecht, Netherlands: Foris. References to the reprint in B. H. Partee (2004), *Compositionality in Formal Semantics: Selected papers by Barbara H. Partee*, pp. 153–181. Oxford: Blackwell.

8

A Guide to Naturalizing Semantics

BARRY LOEWER

Semantic predicates – *is true, refers, is about, has the truth-conditional content that p* – are applicable both to natural-language expressions and to mental states. For example, both the sentence "The cat is crying" and the belief that the cat is crying are about the cat and possess the truth-conditional content that the cat is crying. It is widely thought that the semantic properties of natural-language expressions are derived from the semantic properties of mental states.[1] According to one version of this view, the sentence "The cat is crying" obtains its truth-conditions from conventions governing its use, especially its being used to express the thought that the cat is crying. These conventions are themselves explained in terms of the beliefs, intentions, and so forth of English speakers.[2] In the following I will assume that some such view is correct and concentrate on the semantic properties of mental states.[3]

In virtue of what do mental states possess *their* semantic properties? What makes it the case that a particular mental state is about the cat and has the truth-conditions that the cat is crying? The answer cannot be the same as for natural-language expressions, since the conventions that ground the latter's semantic properties are explained in terms of the semantic properties of mental states. If there is an answer, that is, if semantic properties are real and are not fundamental, then it must be that they are instantiated in virtue of the instantiation of certain non-semantic properties. Recently a number of philosophers, whom I will call "Semantic Naturalizers," have attempted to answer this question in a way that they take to be compatible with *Naturalism*. Naturalism's central contention is that everything there is, every individual, property, law, causal relation, and so on, is ontologically dependent on natural individuals, properties, and so forth. It is not easy or straightforward to spell out the notion of ontological dependence; but for the purposes of this discussion I will understand it as including the claim that for each instantiation of property M there are instantiations of natural properties and relations, P, P*, …, that together with natural laws and causal relations among the P instantiations *metaphysically* entail M's instantiation. This characterization is intended to capture the idea that M is instantiated *in virtue of* the

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

P instantiations. Or, to put it metaphorically, Naturalism is the thesis that for God to create our world He needed only to have created the naturalistic entities and laws. Everything else follows from these.[4]

Naturalists are seldom explicit concerning exactly which properties are the natural ones. Their working account is that the natural properties are those expressed by predicates appropriately definable in terms of predicates that occur in true theories of the natural sciences.[5] Most contemporary naturalists think that all natural-science properties are identical to, or are exemplified in virtue of the exemplification of, fundamental physical properties. These are the properties that occur in laws of fundamental physics. This version of naturalism is physicalism; all God needed to do to create our world was to create the physical properties and laws and set the physical initial conditions. Whether or not they accept physicalism, Semantic Naturalizers assume that certain modal notions, specifically law, causation, and probability, are naturalistically respectable. Whether these notions can be grounded in contemporary physics (or physics and the other natural sciences), or even whether they may presuppose semantic concepts, is not without controversy. Of course, if these notions presuppose semantic notions then they cannot form the basis for a physicalistic or naturalistic reduction of semantics. At best one would have a metaphysical reduction of semantics.[6] Since this issue is seldom addressed by Semantic Naturalizers, and discussing it would involve us in controversial issues in metaphysics, I will, for the most part, ignore it in the following.

Semantic Naturalism is a *metaphysical* doctrine about the status of semantic properties.[7] Semantic Naturalizers also endorse an *epistemic* thesis that I will call "perspicuous semantic naturalism." It is the view that, at least in some cases, the metaphysical connections between naturalistic and semantic properties are sufficiently systematic and transparent to allow us to see that certain naturalistic conditions are sufficient for certain semantic properties. If Semantic Naturalizers were to find naturalistic conditions that are metaphysically sufficient for semantic properties, and know that they have found such conditions, they would show how semantic naturalism can be true and thus place the semantic within the natural order. This guide reviews recent naturalization proposals and the prospects of the naturalization project.

Although Naturalism in something like the above sense is widely endorsed in contemporary philosophy, there is also an active tradition that is inhospitable to semantic naturalism. Adherents to this tradition think that semantic and natural properties are so radically different from each other as to preclude the former from holding in virtue of the latter. Two lines of thought have been especially influential in this regard. One is that semantic properties are essentially normative. A putative example is that it is constitutive of the concept *cat* that it ought to be applied only to cats. Further, it is claimed, such essential normativity cannot be accounted for in purely naturalistic terms.[8] The second line of thought is that the principles that govern the attribution of semantic predicates lead to the indeterminacy of the semantic attributions even given all possible relevant evidence. For example, given all of a person's verbal dispositions (the supposed totality of relevant evidence), principles of attribution license alternative assignments of truth-conditions and references to that person's sentences and terms. It is a verificationist step, but perhaps one that is not inappropriate in this case, to the conclusion that there is no fact of the matter (within the range of indeterminacy) concerning reference and truth-conditions.[9]

There is not a philosophical consensus concerning how far, if any distance at all, these considerations go in undermining semantic naturalism. However, any adequate account of semantic properties will need to account both for the normativity that content properties

possess and for the determinacy of reference and truth-conditions. We will see these issues coming up in various ways in our survey of naturalistic theories.

But first we should note the consequences if semantic naturalism is false. Those who believe that it is false respond in two ways. One is to claim that there are no semantic properties (or that they are never instantiated). This view, Semantic Eliminativism (Churchland, 1981), thus preserves naturalism at the expense of semantics. The other response is to claim that there are semantic properties but they are metaphysically independent of natural properties. This view, Semantic Dualism (Davidson, 1980, especially pp. 207–224 and 245–260; McDowell, 1994), thus preserves semantics at the expense of naturalism. Neither option is very pretty. Eliminativism strikes some philosophers as self-refuting (Boghossian, 1990) and others (Fodor, 1987) merely as obviously false in light of the success of folk-psychological and cognitive-science explanations that employ semantic concepts.[10] Semantic Dualism seems incompatible with semantic properties playing a genuine causal role in producing behavior. If, as is widely believed, the natural sciences are causally complete, then there seems to be no room for causation (of physical effects) in virtue of properties metaphysically independent of natural properties (Papineau, 1993; Loewer, 1995). So the situation seems to be that while there are reasons to worry that semantic naturalism might be false, there are also reasons to doubt the alternatives. The semantic naturalist will resolve this paradox if he can produce a naturalization of semantic properties. That would be enough to quell doubts concerning semantic naturalism, since we would then know that the gap between the semantic and the natural can be bridged.

The mental states that have been the focus of naturalization proposals are the propositional attitudes: desire, belief, and perception (perceptual belief) (see Chapter 14, PROPOSITIONAL ATTITUDES). There are two parts to naturalizing a particular kind of propositional attitude. First is the specifying of natural facts in virtue of which it is an attitude of that particular type, such as a belief or a desire. Second is the specifying of the natural facts in virtue of which it has its semantic properties, such as its particular truth-conditional content. With regard to the first part, the view held by most semantic naturalizers is that the property of being a particular kind of attitude, such as being a belief, is a *functional* property (Fodor, 1987). Functional properties are higher-level properties instantiated by an individual x in virtue of x (or x's parts) and other entities instantiating lower-level properties that are lawfully or causally related to each other in certain specified ways.

Most semantic naturalizers also think that the property of being a belief (or other propositional attitude) involves an internal mental representation, and that this representation bears the state's semantic properties.[11] On this view, for example, the belief that the cat is crying involves a relation to an internal representation that has the truth-conditional content that the cat is crying. Some semantic naturalizers further propose that mental representations are elements in a *language* of thought, "Mentalese."[12] On this view, complex mental representations are composed of names, predicates, logical particles, and so on, arranged in syntactic structures. Naturalizing the semantics of Mentalese consists in specifying the natural facts in virtue of which simple Mentalese expressions possess their semantic properties, and then showing how the semantic properties of complex expressions are determined by their structure and the semantic properties of their constituents (Field, 1972 and 1978). While not every Semantic Naturalizer buys the language of thought hypothesis, it will often be convenient to presuppose it in what follows.

There are two conceptions of semantic content that have figured in recent discussions of naturalizing content, called "broad content" and "narrow content." "Broad content" refers to

the usual truth-conditional content of intentional mental states. Hilary Putnam (1975) posed thought experiments that have been taken to show that the usual truth-conditional content of certain thoughts fails to supervene on the thinker's intrinsic physical properties. Putnam imagined two people, Oscar and twin-Oscar, who are identical with respect to their intrinsic neurophysiological properties, but who differ in the following ways. Oscar lives on Earth and speaks English. Twin-Oscar lives on a twin-Earth and speaks twin-English. The primary difference between Earth and twin-Earth is that on the latter planet the liquid that fills the oceans, that quenches thirst, and so on is not H_2O but XYZ, a chemical compound indistinguishable from H_2O without chemical analysis. Putnam claims that Oscar's and twin-Oscar's utterances of "water is …" and the thoughts that each expresses with the sentence differ in their truth-conditions. Oscar's thought is true iff H_2O is … and twin-Oscar's thought is true iff XYZ is …. If this is correct, then intentional properties, at least in some cases, do not supervene on intrinsic neurophysiological properties or any properties that supervene on them (such as computational or syntactic properties). This view, *semantic externalism*, is now widely held for thoughts that involve natural-kind concepts like *water*.

"Narrow content" is a term introduced to designate content properties that do supervene on intrinsic neurophysiological properties (Fodor, 1981 and 1987). While Oscar and twin-Oscar's thoughts differ in broad content, they agree in narrow content. Some philosophers (Fodor, 1987) have argued that only narrow-content properties are implicated in intentional causation, and for this reason are required by an intentional science; but there is little agreement concerning exactly how to characterize it, or even whether there are such properties (Stalnaker, 1991). In any case, most of the naturalization proposals concern broad properties, specifically reference and truth-conditional content, so that will be our focus here.

What naturalistic facts are plausible candidates to serve as metaphysically sufficient for the semantic properties of mental representations? Putnam's twin-Earth thought experiments and Kripke's well-known theory of proper names (Kripke, 1972) both suggest that causal relations are involved in determining the references of predicates and names (see Chapter 35, REFERENCE AND NECESSITY, §4). Their considerations seem to carry over to mental representations corresponding to predicates and names. It is plausible that Oscar's mental representation "water" refers to H_2O partly in virtue of the fact that H_2O has caused or is apt to cause Oscar to think water thoughts. And it is also plausible that part of the account of what makes a person's mental representation "Aristotle" refer to Aristotle is that it possesses a causal history that originates with a baptism of Aristotle. Neither Putnam nor Kripke are sympathetic to the naturalization project, but their work is often taken as the starting point for naturalistic proposals. Causation and kindred notions like law, counterfactuals, and probability seem to be the "right stuff," if there is right stuff, out of which to try to build naturalistic accounts of intentionality.[13]

The Crude Causal Theory

I will begin our survey of specific naturalization proposals with the crude causal theory (CCT) for the reference of Mentalese predicates f. No one has ever held the CCT, but it will be useful to describe it and note its most obvious defects, since these are the problems that more sophisticated accounts are designed to solve.

(CCT) It is metaphysically necessary that (if tokens of f are caused by and only by instances of the property F then f refers to F).

The obvious problem with the CCT is that it doesn't allow for the possibility of tokening f or a sentence containing f that is not caused by F. This is called "the problem of error," since if f occurs as part of the perceptual belief that x is an f, then since f is caused by F it follows that the belief is true. But of course, a perceptual belief, such as the belief that x is a cat, may be caused by a small dog, not by a cat. The problem of error is a special case of the disjunction problem. The CCT implies that, whether or not f is a component of a belief, the disjunction of all the causes of f's tokens are the reference of f; so if f is caused by cats, small dogs, utterances of "cat," and so on, then CCT says that f refers to the property of being a cat or a small dog or an utterance of "cat," and so on. Clearly many of the causes of f need not be included within what it refers to. A naturalist successor to the CCT will need to find some way of naturalistically distinguishing the reference constituting causes from the others.

A second problem is that semantic relations are apparently more fine-grained than causal relations. This is the "fine-grainedness problem": f may refer to F and not G even though F and G are metaphysically or nomologically co-instantiated. For example, the properties of being triangular and of being trilateral are apparently distinct, but necessarily co-instantiated. Triangular things cause tokens of f just in case trilateral things do, but a predicate can refer to one property but not the other. Quine (1960) pointed out a pervasive type of property co-instantiation. When and only when the property of being a rabbit is instantiated, so is the property of being an undetached rabbit part. When one of these properties is causally linked to f, so is the other. This makes it quite difficult to see how a causal theory can account for the difference between thinking that 'there goes a rabbit' and thinking 'there goes an undetached rabbit part.'

Dretske's Information-Theoretic Account

Fred Dretske (1981) proposed a close relative of the CCT that identifies the truth-conditions of a belief state with part of the information that the state carries under certain circumstances. The notion of information can be defined this way: state type T carries information of type p iff there is a nomological or counterfactual regularity (perhaps a *ceteris paribus* law) to the effect that if a T occurs p obtains.[14] So, for example, the height of mercury in a thermometer carries information about the ambient temperature. Dretske's idea is to construct the content of beliefs out of the information that they carry under certain circumstances. An initial and crude formulation of the theory is:

(DRET) It is metaphysically necessary that (if B carries the information that p then B has the truth-condition that p).[15]

Versions of both the fine-grainedness and the error problem cause trouble for DRET. If B carries the information p and p implies q then it also carries the information that q. But, of course, one can believe that p without believing that q, even though p implies q. Dretske responds to this problem by identifying the content of a belief with the *maximal* information that it carries under certain circumstances. This is a little progress, but it leaves untouched the problem that if p and q are nomologically or metaphysically co-occurring then any state that carries information that p carries the information that q. So according to

DRET, no belief can have the exact content that there is a rabbit, since any state that carries the information that there is a rabbit also carries the information that there is an undetached rabbit part. Notice that it is of no avail to protest that a given believer might not even have the concept *undetached*, since that doesn't affect the fact that his belief state still carries the information that there is an undetached rabbit part. Dretske's attempts to handle this problem are not successful.[16]

The error problem arises for DRET in this way. According to DRET, the belief that p always carries the information that p, which means that whenever the belief is tokened it is true. Dretske's proposal for solving the error problem is to identify a subclass of the actual tokenings of B as the bearers of the information that constitutes B's truth-conditions. Tokens of B outside of this class have the same truth-conditional content as those within the class, although they may not carry the same information. This permits (but doesn't obligate) the latter tokens to be false. Dretske's initial specification of the class of tokens of the belief state that fix its truth-conditional content is the class of tokens that occur and are reinforced during what he calls "the learning period." His idea is that during this period a type of mental state becomes a reliable indicator of p, and so comes to have the content that p. So Dretske's official account is

(DRET*) It is metaphysically necessary that (if the maximal information carried by B during the learning period is p then any instance of B has the truth-condition p).

DRET* allows for errors, but its naturalistic credentials are questionable. The trouble is that *learning* seems to be a semantic notion. Dretske may think that it is possible to characterize the learning period non-semantically, but he can't just take this for granted. In any case, even if the learning period could be characterized naturalistically, the account is implausible, at least for some beliefs. There are some beliefs that are learned in circumstances in which the information they carry is not the belief's content. For example, when a child learns to token a belief with a content about tigers by seeing pictures of tigers, her belief states carry information about pictures, although their content is about tigers. Dretske's account will end up assigning the wrong truth-conditional contents to these beliefs.[17]

Optimal Conditions Accounts

A different way of specifying a belief's content is in terms of the information it would carry under epistemically optimal conditions (Stampe, 1977; Stalnaker, 1984; Fodor, 1990a; 1990b). The core idea of this approach is that there is a class of beliefs for which there are conditions – the epistemically optimal conditions – under which a person has the belief just in case it is true.

(OPT) It is metaphysically necessary that (if B is a belief of kind K then there are epistemically optimal conditions C_B such that B's truth-condition is p if, were C_B the case, then B would nomologically covary with p).

So, for example, if for subject A there is a belief state B, that under optimal conditions covaries with the presence of a red ball located in front of her, then B's content is that there is a red ball in front of A. In this case appropriate optimal conditions are that A's eyes are open, she is attending to what she sees, the lighting is good, and so on.

OPT allows for errors, since tokens of B that don't occur in epistemically optimal conditions need not be true. It also seems to supply truth-conditions with normative force, at least if epistemic optimality is a normative notion. But, like Dretske's theory, its specification of the meaning constituting conditions is not naturalistic. "Epistemically optimal" is clearly an intentional predicate. It is not at all clear that epistemically optimal conditions can be specified without reference to semantic notions. Different conditions are "optimal" for different beliefs. For example, epistemically optimal conditions for the perceptual belief that there is a red ball in the room include good lighting; but optimal conditions for the belief that there is a firefly in the room are that the lights are off. This example makes it obvious that the optimal conditions for acquiring true beliefs depend on the belief's content. Of course the naturalizer cannot appeal to the content of a belief in characterizing optimal conditions.

Not only are epistemically optimal conditions for a belief sensitive to the belief's content, but for most beliefs, if they possess optimal conditions at all, these conditions involve other beliefs. Whether or not a person's belief state reliably covaries with a state of affairs depends on what other beliefs that person has. For example, a person who fails to believe that fossils are derived from once-living organisms, or who believes that the Earth is 6,000 years old, will not reliably form beliefs about the age of a fossil. If there are optimal conditions for forming beliefs concerning the age of fossils, those conditions will involve having certain beliefs and not having certain other beliefs. To assume that optimal conditions can be characterized naturalistically looks as though it begs the naturalization problem rather than solving it.[18]

Teleological Theories

Teleological theories propose to explain the truth-conditional content of mental states, especially certain desires and beliefs, in terms of their biological functions. A crude teleological theory (CTT) for belief is:

(CTT) It is metaphysically necessary that (if O is an organism and B is one of its belief states and it is B's biological function to carry the information that p then B has the truth-conditions that p).[19]

The concept of a biological function is defined in terms of natural selection (Wright, 1973; Neander, 1991) along the following lines: it is the function of biological system S in members of species s to F iff S was selected by natural selection because it Fs.[20] S was selected by natural selection because it Fs just in case S would not have been present (to the extent it is) among members of s had it not increased fitness (that is, the capacity to produce progeny) in the ancestors of members of s.[21] So CTT says that if B was selected because it carried the information that p, then B has the truth-condition that p.

CTT is naturalistic and allows for error. In fact, it is compatible with almost all tokens of B being false, since all that is required is that B was selected because it carried the information that constitutes its content; and that could be so even if most past and no present tokens of B are true. It also seems to supply truth-conditional content with normativity. Just as a heart ought to pump blood, B ought to be tokened only if it carries the information that p. There are, however, a number of problems with CTT. One is that it directly applies only to beliefs composed out of innate concepts, since only beliefs involving innate concepts could possess a biological function. Perhaps the notion of biological function can be extended

beyond features selected by natural selection; but that remains to be seen. A second, and more worrying problem, is that it either fails to assign determinate contents or assigns contents that are much too thick-grained to be the truth-conditions of beliefs. This problem has been discussed mostly with respect to the belief, or proto-belief, of animals, especially a frog's (the hope being that extension to a human's will come when the bugs are worked out).

Suppose that B is an internal state of a frog that is responsive to stimuli and that controls the frog's snapping behavior. Tokens of the state B in the frog's ancestors generally carried a great deal of information including: that flies are present, that small moving black things are present, that food is present, and so on. Furthermore, since these various conditions were reliably co-instantiated in the environment in which the frog evolved, they are all equally good candidates to be the information that it is the function of B to carry. So CTT implies either that B's content is indeterminate among components of the package or that its content is the whole package of information.

It is not clear whether this is an objection to teleological accounts, since it is not clear what beliefs or desires, if any, frogs have. But it is an objection if teleological accounts are incapable of delivering more fine-grained contents than the one they apparently attribute to the frog. More elaborate theories of content that promise to solve this problem are due to Millikan (1984; 1986; 1989) and Papineau (1993). Both accounts, especially Millikan's, are rather elaborate. Here I will just briefly sketch Papineau's approach.

(PAPB) If D is a desire and B a belief and p is the (minimal?) state of affairs whose obtaining guarantees that actions based on B and D satisfy D then B has the truth-condition p.

If we suppose that the frog desires to catch a fly, and that this desire together with B lead to his snapping, then B's truth-conditional content is the minimal state of affairs that will guarantee that snapping will result in catching a fly. In this case it is a belief with something like the content *if I snap then I will catch a fly*. Of course, PAPB is not naturalistic, since it appeals to the concept of satisfying a desire and that is a semantic concept. Papineau attempts to remedy this by providing a naturalistic account of the contents of desires.

(PAPD) If q is the minimal state of affairs such that it is the biological function of D to operate in concert with beliefs to bring about q then D is the desire that q.

Papineau's idea is that if the desire of type D was selected because it contributed by acting in concert with beliefs to bringing about q, then q is D's content. Let's suppose that the content of A's desire D is that she eats an apple. On a particular occasion, D (together with beliefs) may cause the moving of A's hand, A's eating an apple, A's eating a fruit, and A's being nourished. Papineau suggests that the moving of the hand (to grasp the apple) isn't among D's functions, since there are occasions when D was selected (A's ancestors who possessed D had increased fitness, or D was reinforced in A) even though D didn't cause their hands to move. On the other hand, Papineau supposes that whenever D was selected A ate an apple, ate a fruit, was nourished, and so on. He suggests that the most specific of these features of the behavior which led to D's being selected is D's content; that is, eating an apple.

There are a number of worries that one might have concerning Papineau's account. One is that it applies, at best, only to certain beliefs and desires. PAPB provides contents only to means–ends beliefs (although Papineau suggests how the account can be extended to other beliefs). Many desires could not have been selected for by natural selection, since they are

desires that possess impossible satisfaction conditions, or desires for situations that have never obtained, or have obtained too recently to be selected. It is hard to see how the desire to not have any children (or the desire that no one has any children) could have been selected for on the basis of bringing about its content. Perhaps these objections are not all that damaging if PAPD is intended just as a sufficient condition that applies to a certain class of desires. But then we will need a naturalistic specification of that class of desires. More damaging to PAPD is that possessing the function of bringing about x is not a sufficient condition for D's being the desire to bring about x. Suppose that D is the desire to eat an apple. It is compatible with this that there have been occasions when D led not to apple-eating but to pear-eating (some ancestors of A mistook pears for apples). It is plausible that eating pears (pears being as nutritious as apples) led to increased fitness, in which case D's function is to cause (together with beliefs) eating apples or pears. PAPD yields the result, contrary to our assumption, that D is the desire to eat apples or pears. There seems to be no reason why a desire could not have as its function causing, together with beliefs, some situation that differs from its content. If PAPD is incorrect then PAPB, even if it is correct, is no longer adequate as a naturalization of belief.

It is plausible that the human cognitive system contains subsystems that have the functions of producing states that bring about certain effects, and producing other states that carry certain information (and work in concert with the first kind of state to produce effects). But there is no reason to suppose that these states are individuated exactly in the same way that beliefs and desires are. Truth-conditional content seems much more determinate and fine-grained than anything that teleology is capable of delivering. This is made obvious by considering that there cannot be any selectional advantage for creatures whose beliefs are about rabbits over those whose beliefs are about undetached rabbit parts; yet our contents are so fine-grained as to distinguish these belief states.

Fodor's Asymmetric Dependence Theory

Fodor (1990b) proposed a variant of the causal (or informational) account that is intended to be a naturalization of the reference of a simple Mentalese predicate. It appeals to the idea that the meaning-constituting causes are those which, in a sense to be soon explained, are resilient. It will simplify exposition of his theory to define two technical notions. The law $Q \rightarrow C$ (Qs cause Cs) *asymmetrically depends* on the law $P \rightarrow C$ just in case if Ps didn't cause Cs then Qs would not cause Cs but if Qs didn't cause Cs then Ps would still cause Cs. *C locks onto P* just in case (1) it is a law that Ps cause Cs, (2) there are Qs (= Ps) that cause Cs, and (3) for any $Q \neq P$, if Qs cause Cs then Qs causing Cs asymmetrically depends on Ps causing Cs.[22] If C locks onto P then $P \rightarrow C$ is resilient in that it survives the breaking of $Q \rightarrow C$ for Qs other than P. Fodor's proposal, then, is:

(ADT) It is metaphysically necessary that (if C locks onto P then C refers to P).

Suppose that it is a law that cows cause "Cow"s (or rather the word's Mentalese counterpart), that other things also cause "Cow"s, and that such causal relations asymmetrically depend on the 'cow → "Cow"' law. Then, according to ADT, "Cow" refers to cow. ADT handles the error and disjunction problems this way. Horses on a dark night can cause "Cow"s even though the horses on dark nights are not among the reference-constituting causes of "Cow"; that is, the law that horses on a dark night → "Cow"s depends on the law

that cows → "Cow"s. If a horse caused "Cow" is a constituent of the belief "There is a cow," then the belief is false. Of course, this account of error is correct only if ADT is correct. If ADT is not correct then it may count some erroneous beliefs as true, or some true beliefs as erroneous.

Along with the theory, Fodor provides some commentary that helps to understand it. One point is that the law connecting a property to a predicate that refers to it is a *ceteris paribus* law. That is, it holds only as long as certain unspecified conditions obtain. Presumably this means that only under certain kinds of circumstances do cows actually cause A's mental representation "Cow." Presumably these conditions are that cows are perceptually salient to A, A's perceptual system is in good working order, and so on. A second point involves the dependence relation between causal laws. Sometimes Fodor says that it is a basic relation among laws that cannot be explained in other terms. But sometimes he explains it in terms of counterfactuals; Q → C depends on P → C just in case if P → C had not obtained then neither would Q → C have obtained. Fodor insists that the counterfactual be understood *synchronically*, not *diachronically*. If A learned to recognize cows on the basis of pictures of cows, then it may be that cow → "Cow" depends diachronically on cow-picture → "Cow." That is, it is true that if there hadn't been a causal connection between pictures of cows and A's "Cow"s, there wouldn't be a connection between cows and A's "Cow"s. But Fodor thinks that synchronic dependence goes in the opposite direction. Once A has acquired "Cow" then cow → "Cow" is more resilient than cow-picture → "Cow." A third point is that the account of reference is *atomic*. By this is meant that it is metaphysically possible for A's Mentalese predicate C to lock onto P, even if C bears no inferential or causal relations to any of A's other symbols, or even if A's Mentalese vocabulary contains only the predicate C. Fodor welcomes this surprising feature of his account, since he thinks that there are reasons to hold that inferential or causal relations among thoughts are not constitutive of the thought's semantic properties (Fodor and Lepore, 1992).

There are two questions that need answers to evaluate Fodor's theory. First, is it genuinely naturalistic? And, second, is C locking onto P really a sufficient condition for C's referring to P? Answering these questions is made difficult by the fact that the central notions in Fodor's account – *ceteris paribus* laws and asymmetric dependence between laws – are technical notions that are not clearly defined.

There are two places to worry whether ADT is genuinely naturalistic. First, supposing that it is a law that P → C then it is reasonable to believe that its *ceteris paribus* conditions include having and not having certain other intentional states. We noticed a similar point in our discussion of optimal-conditions theories. Does this make P → C non-naturalistic? Not necessarily. If the fact that P → C is a law is naturalistically reducible, then it too is a naturalistic fact. But do we have any reason other than the belief that semantic naturalism is true to think that P → C is naturalistically reducible?

Second, and more worrying, is whether the dependency relations that Fodor requires are naturalistic. These dependency relations are not themselves the subject of any natural science; so Fodor cannot claim, as the teleosemanticist does, that he is explaining a semantic notion in terms of a scientifically respectable notion, that is, a biological function. Further, it is not obvious that the synchronic counterfactuals that Fodor appeals to when explaining asymmetric dependence have truth-conditions that can be specified non-intentionally. Why is Fodor so certain that the counterfactual (synchronically construed) *if cow → "Cow" were broken then cow-picture → "Cow" would also be broken* is true? Perhaps if the first law were to fail "Cow" would change its reference to cow-picture and so the second law would

still obtain. If so, then while "Cow" refers to 'cow,' ADT would say that it refers to 'cow-picture.'[23] Fodor cannot respond by saying that in understanding asymmetric dependence the counterfactual should be understood as holding the actual reference of "Cow" fixed, since that would be introducing a semantic concept into the explanation of asymmetric dependence. I do not think that these points show that ADT is not naturalistic; but they do show that the burden is on Fodor to argue for the naturalistic credentials of the dependency relation. Fodor sometimes seems tempted to just take the dependency relation to be metaphysically primitive and declare that it is part of the natural order (Fodor, 1991). One could see some irony in calling on such elaborate metaphysical notions to defend scientific naturalism.

Is the fact that C locks onto P sufficient for C to refer to P? It is difficult to answer this question without having a clear characterization of asymmetric dependence. The intrepid philosopher who thinks that she has devised a counter-example to ADT runs the risk of being told by its inventor that she has gotten the dependency relations wrong. There are a number of such putative counter-examples in the literature (Baker, 1991; Boghossian, 1991; Adams and Aizawa, 1994; Gates, 1996) and answers to the counter-examples by Fodor (1991; 1994).[24] Instead of going into the details of these objections I will sketch two general worries about the account.

We attribute propositional attitudes to one another on the basis of folk-psychological generalizations and general information about what people tend to believe, desire, and so forth under certain circumstances. So, for example, if A is a normal human being looking at a cow 100 feet away, then we expect A to believe that there is a cow in front of her. If, in fact, there is not a cow but a cleverly made cardboard cow-façade, then we expect A to at first believe that there is a cow, but that when she moves closer to the cardboard cow and examines it she will cease to believe that there is a cow. Our ability to attribute beliefs, desires, and so on to each other depends, at least in part, on generalizations like these. When testing a theory of intentionality we appeal to such generalizations. We ask whether it is possible for the putative naturalistic sufficient condition for A's believing that p to be satisfied while our folk-psychological generalizations give the result that A doesn't believe that p. The problem I see with ADT is not that there are clear cases in which C locks onto P, but C fails to refer to P; it is rather that, as far as I can see, ADT doesn't engage folk psychology. For all we know, an assignment of beliefs to A employing ADT and an assignment employing the usual folk-psychological principles may diverge radically. I am not arguing that they must or do diverge, but that Fodor has provided no reason to think they don't. The worry isn't an idle one, since it is not at all clear what asymmetric dependence has to do with our folk-psychological principles of belief-attribution. If ADT is to carry conviction we need some account of why it is that the contents it assigns will match those assigned by folk psychology.[25]

The second problem is the familiar one of the inscrutability of reference that seems to bedevil all naturalistic theories. If cow → "Cow"s is a law, then so is undetached-cow-part → "Cow" (and laws involving various other properties metaphysically co-instantiated with cow: Quine, 1960). Neither one of these putative laws asymmetrically depends on the other since they hold in exactly the same possible worlds. So it looks like if a predicate locks onto any property it either locks onto all those properties that are metaphysically co-instantiated, or onto the disjunction of all these properties (Gates, 1996).

One response to the problem is to declare that properties like undetached-cow-part, temporal state of a cow, and so on are not eligible to enter into laws and causal relations. Without a naturalistic justification of this claim the response is another instance of borrowing from

metaphysics to buy naturalism. Fodor, to his credit, has not taken this route, but has suggested an addition to ADT to cope with the problem (Fodor, 1994). He argues that the inferential relations among sentences containing the predicate "Cow" will differ (for a thinker whose Mentalese contains the truth-functional connectives) depending on whether "Cow" refers to cow or to undetached cow part. By adding further conditions on the inferential relations borne by sentences to each other, he proposes to specify sufficient conditions for "Cow" to refer to cow (and no other property). The account is too complex to deal with in detail here. I will just say that, at best, Fodor's proposal excludes some properties from being the references of "Cow," but fails to single out cow as the unique reference.

Causal-Role Semantics

Causal-role (aka "conceptual role" and "inferential role") semantics (CRS) is another approach to naturalizing semantics that deserves mention, albeit only a brief one here. The mention is brief because although causal-role semantics has been in the air for some time (Sellars, 1974; Harman, 1982; Field, 1978; Loar, 1981; Block, 1986) no one has actually proposed a CRS that is naturalistic and assigns specific truth-conditions to mental states or representations. The basic idea of CRS is that the semantic properties of a mental representation are partially constituted by certain causal or inferential relations between that and other mental representations. If only causal relations among mental representations are taken into account, then at best CRS is an account of narrow content. To turn it into an account of broad content, causal relations between mental representations and external items need to be added.

CRS should be distinguished from theories of interpretation like Davidson's (1984) that also ground truth-conditions in causal relations among mental representations (or natural-language representations) and external events. Davidson's theory of radical interpretation places constraints on the contents of a person's propositional attitudes. The most important one is that a correct theory of interpretation should assign mostly true beliefs. But the account is not a naturalization, since the semantic concept *truth* is used in formulating the constraint. (On Davidson's theory, see further Chapter 13, RADICAL INTERPRETATION.)

The immediate difficulty with CRS is that most of the actual causal roles of a Mentalese sentence do not seem necessary for it to possess its truth-conditions. For example, a person's Mentalese sentences "There is a cat" and "There is an animal" might have their usual truth-conditions even though the person has no disposition to infer the latter from the former. Given externalism, CRS cannot adequately specify sufficient conditions for a sentence to possess particular truth-conditions solely in terms of its causal connections to other sentences. It will also need to invoke causal connections with external items. But this brings it back to the problem of specifying exactly which causal connections are content-constituting. CRS has made no distinctive contribution to answering this question naturalistically. The prospects for a naturalized CRS do not look good (Fodor and Lepore, 1992).

CRS seems to fare better as an account of what makes it the case that logical expressions possess their meanings. For example, it is plausible that dispositions to infer S from S#R, and to infer S#R from the pair of premises S and R, are relevant to making it the case that "#" is conjunction. But elaborating this into a naturalistic sufficient condition of "#" to be conjunction is not completely straightforward. The most obvious difficulty is characterizing those causal relations that count as *inferences* without appealing to *truth*.

Conclusion

None of the naturalization proposals currently on offer are successful. We have seen a pattern to their failure. Theories that are clearly naturalistic (such as CCT) fail to account for essential features of semantic properties, especially the possibility of error and the fine-grainedness of content. Where these theories are sufficiently explicit we have seen that they are subject to counter-examples. In attempting to avoid counter-examples, semantic naturalists place restrictions on the reference (or truth-condition) constituting causes or information. But in avoiding counter-examples these accounts bring in, either obviously or surreptitiously, semantic and intentional notions, and so fail to be naturalistic.

Of course, the failure of naturalization proposals to date does not mean that a successful naturalization will not be produced tomorrow. But another possibility, and one that philosophers have recently begun to take seriously (such as McGinn, 1993), is that while semantic naturalism is true, we may not be able to discover naturalistic conditions that we can *know* are sufficient for semantic properties; that is, perspicuous semantic naturalism may be false. It may be that the naturalistic conditions that are sufficient for semantic properties are too complicated or too unsystematic for us to be able to see that they are sufficient. Or, it may be that there is something about the nature of semantic concepts that blocks a clear view of how the properties they express can be instantiated in virtue of the instantiation of natural properties. This position, though it may be correct, is not by itself intellectually satisfying. The least we would like to know is exactly why we cannot know which natural properties are sufficient for semantic properties.[26] As of now, we don't know whether semantic naturalism is true and, if it is true, we don't know whether we can know, of any particular proposed naturalization, that it is correct: though, as we have seen, we can know of some that they are incorrect.[27]

Notes

1 Proponents of this view include Grice (1957), Lewis (1969), and Fodor (1975). For a contrary view see Davidson (1984), who holds that mental and public language semantic properties are interdependent, and that neither is metaphysically prior to the other.

2 The program of accounting for the semantic properties of natural language in terms of those of mental states is identified with Paul Grice (1957) and Stephen Schiffer (1972). A detailed account in terms of conventions can be found in Lewis (1969). See also Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING.

3 So in the following, "semantic property" means semantic property of an intentional mental state or event.

4 The proposition that Fx is metaphysically entailed by conditions K just in case K together with a characterization of the nature of F logically imply Fx. The best-understood example of this is the realization of a functional property F by lower-level property instantiations. In this case it logically follows from the functional nature of F, the nature of the Ps, and causal relations among the Ps that whenever the Ps are instantiated M is also instantiated.

5 This characterization is vague with respect to what counts as an appropriate definition, as a property, and as the natural sciences. Removing the vagueness raises a number of problems that would take us too far afield to discuss.

6 Hilary Putnam (see, e.g., 1992) has long maintained that causal and nomological concepts are inextricably bound up with intentionality, and for this reason attempting to naturalize semantics is a misconceived project.

7 Although it is a metaphysical doctrine, it is also contingent, since its truth doesn't rule out possible worlds in which some properties are instantiated but not in virtue of the instantiations of natural properties.

8 There are two issues that are often mentioned by those who think that normativity considerations derail semantic naturalism. One is that grasping a concept involves being in a mental state that obligates one to applying the concept only to items in its extension. It is difficult to see how any purely natural state can entail such an obligation (Kripke, 1982; Boghossian, 1989; see also Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING). The other consideration is the claim that the attribution of intentional concepts is constrained by normative principles of rationality and charity (see Chapter 13, RADICAL INTERPRETATION). Davidson (1980; 1984) starts with this claim and tries to fashion it into an argument against the existence of nomic connections between intentional and non-intentional properties. There is little agreement about exactly what Davidson's argument is or even whether its conclusion conflicts with naturalism. Even so, it has been influential, and is often cited or repeated by those skeptical of naturalization (McDowell, 1994).

9 Quine's (1960) arguments for the indeterminacy of translation and for the inscrutability of reference, and Putnam's (1978) so-called model-theoretic argument are instances of this line of thought (see Chapter 26, INDETERMINACY OF TRANSLATION, and Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM).

10 A sophisticated version of eliminativism maintains that robust semantic properties don't exist (or are uninstantiated) but that deflationary semantic predicates can be used to specify reference and truth-conditions. A robust semantic property is a property that may enter into causal explanations and exists independently of our concepts and definitions. In contrast, a deflationary truth predicate, "DT," for a language L is defined by providing a list of the conditions under which the predicate applies; for example, "Snow is white" is DT iff snow is white; "Snow is green" is DT iff snow is green; etc. More generally (p)("p" is DT iff p) where the quantifier is substitutional. An important feature of DT is that, unlike robust truth, it applies only to the language for which it is defined. There is no reason to suppose that items in the extension of a deflationary predicate have anything, in particular causal and explanatory powers, in common. It seems to follow that deflationary semantic notions cannot be employed in causal explanations or play an explanatory role in an intentional cognitive science. The attraction of deflationism (the view that the only instantiated semantic predicates are deflationary ones) is that it both allows us to use semantic predicates for certain purposes (e.g., for infinite conjunction and disjunction) and is compatible with Naturalism. Skepticism concerning deflationism arises from the worry that deflationary truth and reference are too thin to do the work that we want done by semantic concepts. For discussion see Horwich (1990) and Field (1986; 1994).

11 Proponents of this view usually distinguish between explicit and implicit propositional attitudes. Only the former involve relations to mental representations. The latter are dispositions to produce explicit attitudes (Fodor, 1987, ch. 1).

12 Field (1978) and Fodor (1975; 1987) are important sources of this view. Fodor proposes it as an empirical hypothesis that provides the best explanation of certain features of human thought, specifically systematicity and the capacity to engage in logical reasoning.

13 Causation, laws, counterfactuals, and so on are not themselves items mentioned in physics, and it is controversial whether they supervene on physical facts. Even so, Fodor and other naturalizers would consider it a successful naturalization if they could show that intentional properties supervene on these properties. However, Putnam (1992) has complained that notions of law and causation presuppose intentional notions. While this may be true on some accounts of these notions, it is not true on others. For example, on some accounts, probabilities are rational degrees of belief. Obviously, explaining semantic properties of beliefs in terms of degrees of

belief would not contribute to naturalization. On other accounts, probabilities are objective, mind-independent features of the world. In this case there seems to be no danger of circularity, though one may wonder at employing so metaphysical a notion in the cause of naturalism. However, these issues are too complicated to develop here.

14 Dretske (1981) characterizes information in terms of probabilistic relations. There are numerous problems with his account that are avoided by the characterization used here. Also see Loewer (1987).

15 Dretske's formulation characterizes belief functionally as states that guide behavior in certain ways. He doesn't commit himself to a language of thought account of beliefs.

16 This is forcefully argued in Gates (1996).

17 Dretske (1988) suggests a teleological characterization of the state tokens whose information fixes the beliefs content. His basic idea is that those instances of the belief state that produces behavior that is reinforced are the ones whose informational content fixes the belief's semantic content. While this is a naturalistic characterization of the class, it is questionable whether it assigns appropriate contents. It is easy to imagine situations in which a false token of a belief produces behavior that is reinforced. For further discussion of Dretske's theory see Loewer (1987) and McLaughlin (1993).

18 This point is developed in Loewer (1987) and more thoroughly in Boghossian (1991).

19 Some teleological accounts employ a more general characterization of information. S carries the information that p iff $P(p/S \text{ occurs}) > P(p/S \text{ doesn't occur})$.

20 Selection by conditioning (i.e., by reinforcement) also figures in accounts of function devised by some teleosemanticists (Dretske, 1988).

21 For example, the biological function of the heart is to pump blood (not to make a thumping sound) since it is that property of pumping blood (not making a thumping sound) that accounts via natural selection for the presence of hearts. Notice that something may have the function to F even if it doesn't F or seldom Fs. It should be noted that it doesn't follow that every biological system that does something useful has that as its function (or that it has any function). Only those things that a system does that lead to an increase in fitness are its functions. So, for example, it is not obvious that certain cognitive abilities are the product of any function.

22 Fodor sometimes also adds the requirement that the law $P \rightarrow C$ is instantiated. This is supposed to give the result that Oscar's Mentalese "water" refers to H_2O and twinOscar's Mentalese "water" refers to XYZ. However, this addition may not be needed if the dependency relations concerning laws involving Oscar's and twin-Oscar's mental expressions are different.

23 Boghossian (1991) argues that locking on is either not sufficient for reference or is not naturalistic. His argument shows that to get the counterfactuals that underlie the locking-on relation to come out right, the similarity relation relative to which they are evaluated must take into account *semantic similarities*.

24 One of Boghossian's counter-examples to Fodor's theory is particularly persuasive. He imagines a natural kind concept K and laws $X \rightarrow K$ and $Y \rightarrow K$ where X and Y are different substances that are nomologically indistinguishable by us (they behave differently only in black holes). It may then be that neither of these laws asymmetrically depends on the other. Fodor's theory would have the consequence that K refers to the disjunction $X \vee Y$. But surely in the imagined situation K might refer only to X in virtue of the role it plays in physical theory.

25 This point is developed at length in different ways by Carl Gillett and Andrew Milne in dissertations at Rutgers.

26 Boghossian (1990) argues that belief holism (the fact that which situations are apt to cause one to acquire a particular belief depends on one's other beliefs) prevents us from certifying that any naturalistic condition on content-constituting causes or information is correct.

27 I am grateful to Paul Boghossian, Jerry Fodor, Gary Gates, Carl Gillett, and Fritz Warfield for helpful discussion and (not always heeded) advice.

References

Adams, F., and K. Aizawa 1994. "Fodorian semantics." In *Mental Representation*, edited by S. P. Stich and T. A. Warfield, pp. 223–242. Oxford: Blackwell.

Baker, L. 1991. "Has content been naturalized?" In Loewer and Rey, 1991, pp. 17–32.

Block, N. 1986. "Advertisement for a semantics for psychology." In *Studies in the Philosophy of Mind*, edited by P. French, T. Uehling, and H. Wettstein. *Midwest Studies in Philosophy*, vol. 10. Minneapolis: University of Minnesota Press.

Boghossian, P. 1989. "The rule following considerations." *Mind*, 98(392): 507–549.

Boghossian, P. 1990. "The status of content." *Philosophical Review*, 99(2): 157–184.

Boghossian, P. 1991. "Naturalizing content." In Loewer and Rey, 1991, pp. 65–86.

Churchland, P. 1981. "Eliminative materialism and the propositional attitudes." *Journal of Philosophy*, 78(2): 67–90.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.

Field, H. 1972. "Tarski's theory of truth." *Journal of Philosophy*, 69(13): 347–375.

Field, H. 1978. "Mental representation." *Erkenntnis*, 13: 9–61.

Field, H. 1986. "The deflationary conception of truth." In *Fact, Science, and Value*, edited by G. McDonald and C. Wright, pp. 55–117. Oxford: Blackwell.

Field, H. 1994. "Deflationist views of meaning and content." *Mind*, 103(411): 249–285.

Fodor, J. 1975. *The Language of Thought*. New York: Thomas Y. Cromwell.

Fodor, J. 1981. "Methodological solipsism." In *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Brighton: Harvester.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1990a. "Psychosemantics, or where do truth conditions come from." In *Mind and Cognition*, edited by W. Lycan, pp. 312–338. Oxford: Blackwell.

Fodor, J. 1990b. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1991. "Replies." In Loewer and Rey, 1991, pp. 255–319.

Fodor, J. 1994. *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, J., and E. Lepore. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.

Gates, G. 1996. "The price of information." *Synthese* 107(3): 325–347.

Gillett, C. 1997. "Naturalization: Physicalism and Scientific Theory Appraisal." PhD diss., Rutgers University.

Grice, P. 1957. "Meaning." *Philosophical Review*, 66(3): 377–388.

Harman, G. 1982. "Conceptual role semantics." *Notre Dame Journal of Formal Logic*, 23(2): 242–256.

Horwich, P. 1990. *Truth*. Oxford: Blackwell.

Kripke, S. 1972. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Lewis, D. 1969. *Convention*. Cambridge, MA: Harvard University Press.

Loar, B. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.

Loewer, B. 1987. "From information to intentionality." *Synthese*, 70(2): 287–317. Reprinted in Stich and Warfield, 1994.

Loewer, B. 1995. "An argument for strong supervenience." In *New Essays on Supervenience*, edited by E. Savellos, pp. 218–225. Cambridge: Cambridge University Press.

Loewer, B., and G. Rey, eds. 1991. *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

McDowell, J. 1994. *The Mind and the World*. Cambridge, MA: Harvard University Press.

McGinn, C. 1993. *Problems in Philosophy: The Limits of Inquiry*. Oxford: Blackwell.

McLaughlin, B., ed. 1993. *Dretske and his Critics*. Oxford: Blackwell.
Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Millikan, R. 1986. "Thoughts without laws: cognitive science with content." *Philosophical Review*, 95(1): 47–80.
Millikan, R. 1989. "Biosemantics." *Journal of Philosophy*, 86(6): 281–297. Reprinted in Stich and Warfield, 1994, pp. 243–258.
Milne, A. 1996. "The Alienation of Content: Truth, Rationality and Mind." PhD diss., Rutgers University.
Neander, K. 1991. "Functions as selected effects: the conceptual analyst's defense." *Philosophy of Science*, 58(2): 169–184.
Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.
Putnam, H. 1975. *Mind, Language and Reality (Philosophical Papers,* vol. 2*)*. Cambridge: Cambridge University Press.
Putnam, H. 1978. *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul.
Putnam, H. 1992. *Renewing Philosophy*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Schiffer, S. 1972. *Meaning*. Oxford: Oxford University Press.
Sellars, W. 1974. "Meaning as functional classification." *Synthese*, 27(3–4): 417–437.
Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, R. 1991. "Semantics for the language of thought." In Loewer and Rey, 1991, pp. 229–237.
Stampe, D. 1977. "Towards a theory of linguistic representation." *Midwest Studies in Philosophy*, 2(1): 42–63.
Stich, S., and T. Warfield, eds. 1994. *Mental Representation*. Oxford: Blackwell.
Wright, L. 1973. "Functions." *Philosophical Review*, 84(2): 139–168.

Further Reading

Field, H. 1977. "Logic, meaning, and conceptual role." *Journal of Philosophy*, 74(7): 379–409.
Kim, J. 1993. *Supervenience and Mind*. Cambridge: Cambridge University Press.
McGinn, C. 1982. "The structure of content." In *Thought and Object*, edited by A. Woodfield, pp. 207–259. Oxford: Oxford University Press.
Millikan, R. 1991. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.
Neander, K. 1995. "Misrepresenting & malfunctioning." *Philosophical Studies*, 79(2): 109–141.
Pietroski, P. 1992. "Intentionality and teleological error." *Pacific Philosophical Quarterly*, 73: 267–282.

Postscript

PETER SCHULTE

How can mental states be about things in the world? How is it possible that certain mental states are true (or satisfied) under some conditions, and false (or unsatisfied) under others? These questions continue to puzzle philosophers, especially those of a naturalistic bent, since all attempts to explain semantic properties in naturalistic terms face serious difficulties. This has led some theorists to conclude that the whole naturalization program is wrongheaded, but others still adhere to it: while admitting that the task of naturalizing semantic properties is more difficult than naturalists might have initially supposed, they see no reason to be pessimistic about the program as a whole. Hence, the debate about naturalizing semantics – or, more specifically, about naturalizing mental content – continues, and has even intensified in recent years. Since teleological theories are at the center of this debate, I concentrate on them in what follows.

Teleological Theories: Basic Distinctions

It is useful to distinguish between two questions which semantic naturalizers must address (Sterelny, 1995, p. 254). The first question concerns *representational status*: What distinguishes representational states, that is, states with semantic content, from non-representational states? Or, more precisely, in virtue of which natural facts do certain internal states qualify as representational states? Call this the 'status question.' The second question concerns *content determination*: Given that R is a representational state, why does R have the content that p rather than some other content? To put it differently, which are the natural facts that determine the content of R? Call this the 'content question.' Of course, those two questions are intimately related, and many theorists answer them both at once (e.g., Millikan, 1984), but it is important to emphasize that they can be discussed separately.

Traditionally, semantic naturalizers have focused on the content question, and they have continued to do so in recent years. One of the main debates in this area is the debate between proponents of two different teleological approaches to content determination, the *input-* and the *output-oriented* approach. These approaches differ fundamentally in the way they specify the content of descriptive representational states (i.e., 'belief-like' states with a mind–world direction of fit). Consider the descriptive representational state R. According to the input-oriented approach, we have to look 'upstream' at the functions of the mechanisms that generate R, or to R's functional relations to causes or conditions in the world, in order to specify the semantic content of R. One example of an input-oriented account is the "crude teleological theory" (p. 180), which equates R's semantic content with the information that R is supposed to carry (Dretske, 1988). According to the output-oriented approach, on the other hand, R's content is primarily dependent on 'downstream factors' – for example, on the biological function of mechanisms that respond to R, or on R's functional relations to behavior. Papineau's (1993; 1998) teleological theory is a variant of the latter approach, since he proposes that belief content depends on desire content, and that the content of a desire is the most specific state of affairs it is supposed to bring about (p. 181).

The Content Question: Input-Oriented Theories

An attractive new version of the input-oriented approach is Karen Neander's (2013) causal-informational version of teleosemantics. It should be noted that her theory is restricted to a subclass of descriptive representational states – namely, to perceptions. According to Neander, the content of a perceptual state is determined by a certain function of the perceptual mechanism that produces it (i.e., by a function of its 'producer'). More precisely, it is determined by the producer's *response function*: the function to generate R in response to certain environmental conditions p. So Neander's teleological theory is

(NTT) If R is a perceptual state and R's producer has the function of producing R in response to the state of affairs p, then R has the content that p.

Consider again the case of the frog (p. 181), and assume that N is the neural state that (a) is normally produced when there's a small, dark, moving object in the frog's visual field, and (b) normally triggers prey-catching behavior. According to Neander's account, N has the

content that a small, dark, moving object is present, because the frog's visual system has the function of producing N in response to a small, dark, moving object (cf. Neander, 2013, p. 31).[1] *Responding to a small, dark, moving object by producing N* is what, among other things, the frog's visual system has been selected for.

At this point, one might ask: Can't we also ascribe the function of *responding to a nutritious object ('frog food') by producing N* to the frog's visual system? What's more, isn't this description in some sense superior, given that it is only because small, dark, moving objects often contained nutrients that the system was selected for responding to them? This seems to raise the notorious indeterminacy problem all over again (p. 181). However, it should be noted that Neander uses the term 'responding' in a strictly causal sense (Neander, 2013, p. 23). A mechanism can only be selected for responding to some x's being F by producing R if x's being F is *causally relevant* for bringing about the mechanism's production of R. But in the case of the frog, the object's being nutritious is *not* causally relevant for bringing about the visual system's production of N, only the object's being small, dark, and moving is (as can easily be tested by varying the properties independently). Hence, the system can only have been selected for responding to small, dark, moving objects. This is true even though the evolutionary *reason* why the system was selected for doing this consists in the fact that small, dark, moving objects were (often enough) nutritious. (Since Neander also analyzes information in terms of causation, she argues that NTT can be restated as a version of informational teleosemantics. This claim, however, is not essential to her theory.)

Neander's input-oriented account of perceptual content is thus of great interest, since it promises to solve the indeterminacy problem that was fatal to earlier input-oriented theories. Still, there are several objections that can be raised against it. One worry concerns the notion of a response function (Millikan, 2013). Another worry centers on a second problem of indeterminacy, often called the 'distality problem.' This problem becomes apparent when we consider that the frog's N-state normally stands at the end of a long chain of causes – it is caused by a pattern of retinal stimulation, which is caused by a pattern of light waves, which in turn is caused by the presence of a small, dark, moving object. So which of these 'normal causes' constitutes the content of N, according to NTT? Appealing to the visual system's response functions is of no help here: the system has the function f_1 to produce N in response to a certain pattern of retinal stimulation, but also the function f_2 to produce N in response to a certain pattern of light and the function f_3 to produce N in response to a small, dark, moving object (it performs f_3 *by* performing f_2, and f_2 *by* performing f_1). Thus the content of N turns out to be indeterminate. Neander (2013, pp. 33–35) tries to solve this problem by slightly modifying NTT, but it remains to be seen whether this modified account really yields plausible content ascriptions for all relevant cases.

The Content Question: Output-Oriented and Mixed Theories

The most discussed version of the output-oriented approach, and the most discussed version of teleosemantics generally, is Millikan's biosemantics. Millikan first presented her theory in *Language, Thought, and Other Biological Categories* (1984), but has extended and refined it ever since (cf. Millikan, 1989; 2004; 2009). According to Millikan, representations are basically signals passing from a producing mechanism ('producer' or 'sender') to a consuming mechanism ('consumer' or 'receiver'), and descriptive representations are those signals which have to vary systematically with conditions in the environment in order for

the consumer to be able to fulfill its functions. More formally, but still simplified, Millikan's teleological theory can be summarized as follows:

(MTT) R_i is a descriptive representation with the content that p iff (i) R_i belongs to a family of states R_1, …, R_n (the 'R-signals') which stand midway between two cooperating devices, a producer and a consumer, (ii) for the consumer to fulfill its biological functions in a normal way, different R-signals must correspond systematically to different conditions in the external world, and (iii) R_i must correspond to condition p.[2]

To determine the content of a descriptive representation, it is thus crucial to look at the biological functions of the *consumer* of that representation. Since the consumer is located 'downstream' from R, this makes MTT into a version of the output-oriented approach.

The difference between MTT and an input-oriented theory like NTT becomes clearer when we consider what MTT says about the frog case. Here, the relevant family of representations consists entirely of (possible) N-tokens occurring at different times: <N, t_1>, …, <N, t_n>. The producer of these states is the frog's visual system, and their consumer is the prey-catching mechanism – the mechanism responsible for the frog's snapping behavior. The main function of this mechanism is to catch prey, and thus ultimately to provide the organism with nutrients. Since the mechanism is activated by N-states, a covariation between N-states and small (i.e., swallowable) nutritious objects is necessary for the mechanism to fulfill its functions, that is, the representational states < N, t_1>, …, <N, t_n > must correspond to the conditions < small nutritious object present, t_1>, …, <small nutritious object present, t_n>. By contrast, it is irrelevant for the consumer's well-functioning whether the object in front of the frog is dark or moving: the prey-catching mechanism, once activated by N, would fulfill its functions equally well if the small nutritious object were light and stationary. (Of course, it would normally not *get* activated under these conditions, because N would not occur, but from Millikan's output-oriented perspective, this makes no difference to N's content.)

The same point can be put another way. In the evolutionary past, earlier tokens of the prey-catching mechanism have often provided their possessors with objects that were small, dark, and moving. Many of those objects were flies, and most of the flies were nutritious. But according to Millikan, it is only because the prey-catching mechanism provided its possessor with (small) *nutritious* objects that it was evolutionarily successful, so it acquired only the function of catching those objects, not the function of catching small, dark, moving objects or flies. Hence, MTT implies that an N-token occurring at t_1 has the content that there is a small nutritious object present at t_1 (cf. Millikan, 1991, p. 163).

Many objections have been raised against Millikan's proposal. A very influential criticism stems from Neander (1995, pp. 126–127). She argues that MTT, rigorously applied, yields content ascriptions that are *overly specific*. When we take a closer look at the frog case, for example, we find that the fact that really explains the evolutionary success of the prey-catching mechanism is not the fact that earlier tokens of the mechanism provided their possessors (often enough) with small nutritious objects *tout court*, but rather the fact they provided them with small objects that contained enough nutrients to make up for the calories lost in catching and digesting them, and that were, in addition, digestible, free from poison, not contaminated with deadly pathogens, and so on. So the mechanism must have the *function* to provide the organism with objects that have all these properties, and this entails, according to MTT, that all these properties enter into the content of N. To be sure, these consequences are highly implausible.

In recent work, Millikan has replied to Neander's objection by bringing in the functions of representation *producers* and the normal mechanisms by which they perform these functions (Millikan, 2004, pp. 85–86; 2009, p. 404), but it is not entirely clear how this reply is supposed to work in detail and whether it is consistent with Millikan's official output-oriented approach.

Even if the problem of overly specific contents can be avoided, another question remains: Does MTT entail content ascriptions that are *plausible*? Pietroski (1992) construes a case of hypothetical creatures called 'kimus' and argues that the content ascriptions entailed by MTT for this case are in conflict with our pre-theoretic intuitions. Defenders of MTT, however, can simply reject those intuitions as irrelevant (cf. Millikan, 2009, pp. 405–406). Consequently, Neander (2006) pursues a different strategy to attack the plausibility of the content ascriptions entailed by MTT. She considers the states that govern prey-catching behavior in toads, and argues that the content ascribed to these states by MTT (namely that there is a small nutritious object present at t_1, or something along these lines) is implausible from the perspective of mainstream cognitive science, since this content ascription does not fit with standard information-processing explanations of the toad's discriminatory capacities.

Not every teleological account of content can be neatly categorized as a version of either the input-oriented or the output-oriented approach. Some theorists defend mixed theories, holding that 'upstream' *and* 'downstream' factors are important for determining the content of descriptive representations. Carolyn Price (2001, pp. 89–103) and Nicholas Shea (2007, p. 418), for example, add to a broadly Millikanian theory of content the requirement that R must carry *information* about p in order to have the content that p. The addition of an input-requirement of this kind may help to rule out overly specific or otherwise implausible contents: one could argue, for example, that the frog's N-state does not carry information about the absence of pathogens in frog food, so that *not being contaminated with pathogens* cannot enter into N's content. Whether this strategy is successful depends, however, on the details of the theory, and especially on the definition of 'information' that is employed.

The Status Question

So far, the focus of this postscript has been on teleosemantic answers to the content question. But it is important to note that some theories also provide answers to the status question. Most prominently, Millikan's MTT specifies the conditions that are necessary and sufficient for conferring the status of a descriptive representation on R. These conditions include, besides clauses (ii) and (iii) which are also crucial for determining R's precise content, the condition that R must stand midway between two mechanisms, a producer and a consumer, which are designed to cooperate with each other (clause (i)). Thus, at least when it comes to representational status, Millikan's theory is not purely consumer- or output-oriented.[3]

Millikan's answer to the status question has been criticized in several different ways. Some theorists argue that MTT is *too restrictive*, because it requires representations to have cooperating producers and consumers (Sterelny, 1995; Stegmann, 2009). This requirement appears to exclude non-cooperative animal signals like the 'stotting' of gazelles, a behavioral display which (*prima facie*) indicates to approaching predators that they have been recognized, and this may seem problematic (for a reply on Millikan's behalf, see Artiga, 2014). Stegmann (2009) uses similar cases to motivate a purely output-oriented version of teleosemantics, where representational content *and* representational status are determined exclusively by consumer mechanisms and their functions.

Most critics, however, argue that Millikan's status requirements are *too liberal*. Price (2001, pp. 93–96) and Shea (2007, pp. 427–430) describe (hypothetical) organisms whose feeding behavior is triggered by a randomly generated internal state. Since food is abundant in the habitat of these organisms, this way of producing feeding behavior is evolutionarily successful. Millikan seems to be committed to treat the randomly generated internal state of such an organism as a descriptive representation of food, which does not seem adequate. Price and Shea suggest that this defect can be corrected by adding an informational input-requirement to MTT (see above).

But Millikan's status requirements may also be too liberal in a more fundamental way. Burge (2010) argues that Millikan's theory describes many simple systems as representational even though this description yields no explanatory benefits, thus drawing the 'lower border of representation' too low. He even suggests that this defect is shared by all teleological accounts of representation, but this assessment is arguably due to an impoverished conception of the resources available to the teleological approach (Schulte, 2015).

Conclusion

Explaining content in a naturalistic way is not an easy task, and for all we know, it might turn out to be impossible. But the naturalistic proposals that have been formulated and defended in recent years surely deserve serious consideration, and should not be dismissed out of hand. This goes for the theories presented above as well as for those omitted here like, for example, the teleological accounts of Dan Ryder (2004), Mohan Matthen (2005), and Manolo Martínez (2013), or the non-teleological account of Robert Rupert (1999).

Alternatives to Naturalized Semantics have also found new supporters in the past few years. Some adherents of the phenomenal intentionality paradigm accept that semantic content is a primitive feature of mental states, not capable of naturalistic explanation (e.g., Strawson, 2008), and proponents of radical embodied or enactive approaches to cognition deny that there are any contentful mental states at all (e.g., Noë, 2009). But neither of these alternatives has gained wide acceptance among contemporary philosophers. So if one thing is clear, it is that the discussion about Naturalized Semantics will continue for many years to come.[4]

Notes

1 Here and in the following, I'm disregarding a number of empirical details about the case, for example, the fact that the frog's perceptual states also represent the *location* of the object in the visual field.

2 Millikan (1984, pp. 96–97; 2009) spells out the last two conditions in terms of "mapping functions" or "rules of correspondence," but to introduce this terminology here would complicate the matter unnecessarily.

3 For a different take on the status question, see Papineau (2003).

4 I would like to thank Hannah Altehenger, Fabian Hundertmark, Insa Lawler, and Alex Miller for helpful comments.

References

Artiga, M. 2014. "Signaling without cooperation." *Biology & Philosophy*, 29(3): 357–378.

Burge, T. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.

Martínez, M. 2013. "Teleosemantics and indeterminacy." *Dialectica*, 67(4): 427–453.

Matthen, M. 2005. *Seeing, Doing, and Knowing*. Oxford: Oxford University Press.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. 1989. "Biosemantics." *Journal of Philosophy*, 86(6): 281–297. Reprinted in Stich and Warfield, 1994.

Millikan, R. 1991. "Speaking up for Darwin." In *Meaning in Mind: Fodor and His Critics*, edited by B. Loewer and G. Rey, 151–164. Oxford: Blackwell.

Millikan, R. 2004. *Varieties of Meaning*. Cambridge, MA: MIT Press.

Millikan, R. 2009. "Biosemantics." In *The Oxford Handbook of Philosophy of Mind*, edited by B. McLaughlin, A. Beckermann, and S. Walter, pp. 394–406. Oxford: Oxford University Press.

Millikan, R. 2013. "Reply to Neander." In *Millikan and Her Critics*, edited by D. Ryder, J. Kingsbury, and K. Williford, pp. 37–40. Chichester: Wiley-Blackwell.

Neander, K. 1995. "Misrepresenting & malfunctioning." *Philosophical Studies*, 79(2): 109–141.

Neander, K. 2006. "Content for cognitive science." In *Teleosemantics*, edited by G. MacDonald, and D. Papineau, pp. 167–194. Oxford: Oxford University Press.

Neander, K. 2013. "Toward an informational teleosemantics." In *Millikan and Her Critics*, edited by D. Ryder, J. Kingsbury, and K. Williford, pp. 21–36. Chichester: Wiley-Blackwell.

Noë, A. 2009. *Out of Our Heads*. New York: Hill and Wang.

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

Papineau, D. 1998. "Teleosemantics and indeterminacy." *Australasian Journal of Philosophy*, 76(1): 1–14.

Papineau, D. 2003. "Is representation rife?" *Ratio*, 16(2): 107–123.

Pietroski, P. 1992. "Intentionality and teleological error." *Pacific Philosophical Quarterly*, 73: 267–282.

Price, C. 2001. *Functions in Mind: A Theory of Intentional Content*. Oxford: Oxford University Press.

Ryder, D. 2004. "SINBAD neurosemantics: a theory of mental representation." *Mind & Language*, 19(2): 211–240.

Rupert, R. 1999. "The best theory of extension: first principle(s)." *Mind & Language*, 14(3): 321–355.

Schulte, P. 2015. "Perceptual representations: a teleosemantic answer to the breadth-of-application problem." *Biology & Philosophy*, 30(1): 119–136.

Shea, N. 2007. "Consumers need information: supplementing teleosemantics with an input condition." *Philosophy and Phenomenological Research*, 75(2): 404–435.

Stegmann, U. 2009. "A consumer-based teleosemantics for animal signals." *Philosophy of Science*, 76(5): 864–875.

Sterelny, K. 1995. "Basic minds." *Philosophical Perspectives*, 9: 251–270.

Stich, S., and T. Warfield, eds. 1994. *Mental Representation*. Oxford: Blackwell.

Strawson, G. 2008. "Real intentionality 3: why intentionality entails consciousness." In *Real Materialism and Other Essays*, pp. 281–305. Oxford: Clarendon Press.

9

Inferentialism

JULIEN MURZI AND FLORIAN STEINBERGER[1]

There are two *prima facie* opposing tendencies in philosophical work concerning linguistic meaning and mental content: 'referential' approaches and 'use-theoretic' approaches. According to referential approaches the semantic properties of linguistic expressions and concepts are primarily to be explained in terms of the (broadly) referential relations they bear to (typically) extra-linguistic things (objects, sets thereof, instantiations of properties, etc.). In the case of language, referential relations are understood in terms of mappings from linguistic expressions to the corresponding semantic values in accordance with their semantic types: proper names are mapped onto the appropriate objects, simple predicates might be associated with the properties they designate, and so on. Our use of linguistic expressions is then explained *in terms* of these referential semantic properties: we use the expressions of our language as we do because of their referential properties. A roughly analogous story, it is generally thought, can be told about thought. Use theories of meaning, in contrast, reverse the order of explanation. According to them, it is regularities or rules of use that take center stage. It is these regularities and rules that are accorded explanatory primacy in accounting for meaning and conceptual content, and customary semantic notions such as reference, truth, and satisfaction are explained as a by-product of them. Thus, linguistic practice precedes and shapes semantic theory rather than the other way around.

Plausibly, on a use-theoretic approach, not all aspects of an expression's or a concept's use are equally explanatorily relevant. Accordingly, use theories of meaning differ over which features of an expression's use does most of the explanatory work. Our focus in this chapter is with the particular class of use theories that gives *inference* pride of place in their account of meaning. Use theories of meaning of this sort are commonly known as *inferential role semantics* (IRS) or *inferentialism* for short. IRS is a species of *conceptual role semantics* (CRS). On a broad understanding of it, conceptual role semantics includes "any theory that holds that the content of mental states or symbols is determined by any part

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

of their role or use in thought" (Greenberg and Harman, 2006, p. 295). IRS, as we conceive of it, restricts the semantically relevant features of an expression's conceptual role to the regularities or proprieties of inference (or to some particular subclass of these). The meaning of 'and,' for instance, is often said to be determined by the rules of inference governing it. And to understand 'and,' that is, to know what 'and' means, is to infer according to such rules.

In this chapter, we introduce IRS and some of the challenges it faces. We aim to provide a map of the terrain which offers an overview, but also challenges some of the inferentialist's standard commitments. Our discussion is structured thus. §1 introduces inferentialism and places it into the wider context of contemporary philosophy of language. §2 focuses on what is standardly considered both the most important test case for and the most natural application of IRS: logical inferentialism, the view that the meanings of the logical expressions are fully determined by the basic rules for their correct use, and that to understand a logical expression is to use it in accordance with the appropriate rules. We discuss some of the (alleged) benefits of logical inferentialism, chiefly with regard to the epistemology of logic, and consider a number of objections. §3 introduces and critically examines Robert Brandom's inferentialism about linguistic and conceptual content in general. Finally, in §4 we consider a number of general objections to IRS and consider possible responses on the inferentialist's behalf.

1 Varieties of Inferentialism

Inferentialism is a broad church. In this section, we introduce what we take to be its main varieties, and place them in the wider context of contemporary philosophy of language.

To begin, IRS might be thought to serve as a theory of linguistic meaning. As such, it might be thought of as a theory of the meanings of expressions in a *public* language. Alternatively, it might be taken to be an account of the meanings of expressions in someone's idiolect, or as an account of the contents of symbols in a language of thought, or again it might be thought of as an account of the content of thought.

Some clarification concerning the intended sense of 'thought' is in order. Thoughts, as we will understand them, are the sorts of things that can be grasped or entertained. Thoughts are commonly thought to be composed of concepts displaying something resembling syntactic structure. It follows that thoughts so understood differ from theories that construe propositions as 'unstructured,' for example theories that identify propositions with sets of possible worlds. Thoughts differ also from Russellian propositions, which are composed of the objects, properties, and relations they are about. As Timothy Williamson puts it, "a thought about Vienna contains the concept of Vienna, not Vienna itself" (Williamson, 2006, p. 2). If Fregean accounts of propositions are correct, then thoughts may be propositions. If they are not, thoughts may still express propositions. Though we will not take a stand on these issues here, laying out the options and situating our talk of meaning and conceptual content within them should clarify the discussion to follow.

It is a further question how certain aspects of either of the three aforementioned kinds of meaning – meanings of public language expressions, meanings of idiolectal expressions, and meanings of expressions of a language of thought – relate to conceptual content. Some advocates of IRS maintain that mastery of public linguistic meaning is conceptually prior to conceptual content; that one can only have concepts (or at least concepts of

any degree of sophistication) once one has grasped the meanings of the corresponding linguistic expressions by knowing how to use them (Dummett, 1991; 1993; Sellars, 1956). Others hold the weaker view that linguistic meaning is methodologically prior to conceptual content, or at least that thought and talk must go hand-in-hand in that they must be accounted for in unison (Brandom, 1994; Davidson, 1984; Harman, 1999). Thus, on most forms of IRS, the only way to account for conceptual content is by way of an account of linguistic meaning. The thought, usually, is that concepts can only be attributed to creatures capable of manifesting them, and that linguistic competence is required in order to do so. But that is not to say that all inferentialist accounts abide by this language-first methodology. Loar (1981) and Peacocke (1992) both advance views whereupon inferential roles determine mental content, but where linguistic meaning is parasitic on mental content. Linguistic meaning might then be derived from mental contents according to roughly Gricean lines (Grice, 1957). With this point being noted, in the following we will allow ourselves to slide from (linguistic) meaning talk to (conceptual) content talk.

We have presented referential semantics and use-theoretic semantics (and IRS in particular) as two '*prima facie* opposing' approaches to meaning and conceptual content. Treating them as opposing alternatives is common practice and indeed partisans on either side often proclaim their opposition to the other side. However, as we have also noted, the difference between IRS and referentialism need not necessarily reveal itself at the level of semantic theory; rather, it is a difference *in the order of explanation*: 'Are (broadly) referential relations explanatorily prior to inferential ones or does the order of explanation run in the opposite direction?' And *this* question is in fact a *meta-semantic* question, not a semantic one. The underlying distinction at play here comes to this. *Semantics* concerns itself with the question of which types of semantic values to assign to different categories of expressions, and how, on the basis of these assignments, the semantic values of complex expressions are functionally determined by the semantic values of their constituent expressions. Meta-semantics,[2] in contrast, asks two questions: the *metaphysical* question as to what makes it the case that a given expression has the semantic value it does; and the *epistemological* question as to what a speaker needs to know to count as understanding the expression. Whence our claim that IRS is *not* incompatible with a referential *semantics*. It is incompatible with the meta-semantic thesis that referential relations precede inferential ones in the order of explanation.

A proponent of IRS could thus in principle adhere to a referential *semantic* theory – that is, one that deals in the ordinary semantic concepts of reference, truth, and satisfaction and which assigns the customary (typically) extra-linguistic items as semantic values, while at the same time staying true to the spirit of IRS. IRS might then be interpreted as a meta-semantic thesis – a thesis about what it is in virtue of which an expression has the semantic value it does (what it is in virtue of which a concept has the content it does) or a thesis about what it takes to understand an expression (grasp a concept), or both. IRS, on this meta-semantic interpretation, gives rise to the following two broad theses:

(MD) *Meaning determination*. The meanings of linguistic expressions are determined by their role in inference.

(UND) *Understanding*. To understand a linguistic expression is to know its role in inference.[3]

Thus, a position that takes referential semantics at face value and appeals to IRS as a source for answers to meta-semantic questions is at least conceivable. That being said, given the explanatory priority IRS accords to inferential over referential relations, IRS paves the way for non-standard semantic theories. This may be either because it is thought that IRS constrains semantic theory in such a way as to necessitate the assignment of non-standard semantic values (e.g., assertibility conditions instead of truth-conditions or an epistemically constrained notion of truth in the case of anti-realist theories of meaning like those of Michael Dummett, Crispin Wright, etc.); or it may be because it is felt that substantive relations of reference and truth are not warranted or perhaps not needed once an IRS account of meaning has been offered.[4] Let us call IRS accounts coupled with 'standard' semantic theories *orthodox*, and *unorthodox* otherwise. Moreover, let us call *genuine* those orthodox accounts that take the semantic concepts featured in the semantic theory at face value, and orthodox accounts that are not genuine, *deflationary*.

To bring out some of the characteristic features common to all forms of IRS as well as some of the features by which different of its variants distinguish themselves from one another, it will be helpful to contrast IRS with a theory that *is* straightforwardly incompatible with it. Now, as a species of use-theoretic account, IRS is at odds with any account that takes expressions of a language or concepts to come, as it were, pre-equipped with meanings or contents. Views that run under the banner of informational semantics are a case in point. Roughly, advocates of informational semantics maintain that semantic concepts are to be explained in terms of certain lawlike correlations linking external things or property instantiations to tokenings of corresponding linguistic items or to mental items (Fodor, 1990; Dretske, 2000). The primary mode of semantic explanation thus proceeds by establishing 'direct' language–world mappings. Once the reference and designation relations are established by way of the said reliable correlations between linguistic or mental items and their external causal antecedents, the usual semantic concepts (reference, truth, satisfaction, etc.) are simply explained from the 'bottom up': first, names are associated with their bearers, unary predicates with the properties designated, and so on. In a second step, the theory then specifies compositional rules for determining the semantic values of more complex expressions as a function of their semantically relevant component parts. This atomistic mode of explanation is more congenial to certain types of expressions than it is to others. Paradigmatic cases of linguistic expressions that particularly lend themselves to informational semantic explanation are observational predicates ('square,' 'red') and proper names.

If the prototypical expressions for informational semantical treatment are observational predicates and proper names, the paradigm case for IRS are logical expressions ('and,' 'or,' 'if,' etc.). The semantic values of logical expressions are not readily explained in terms of correlations that may obtain between tokenings of them and external referential relata. Rather, the inferentialist will maintain, the meanings of logical expressions are determined, first and foremost, by their inferential properties which are encapsulated in the rules governing paradigmatic inferences involving them. Taking the case of logical expressions as its model, IRS takes the bulk of the explanatory work to be done by *intra*-linguistic (or language–language) relations (as opposed to by direct language–world links). Moreover, while informational semantics and kindred approaches proceed from the *bottom up* as we have seen, the inferentialist mode of explanation is *top down*: simple declarative sentences are the primary semantic units,

for it is the propositions expressed by them that stand in inferential relations. The meanings of sub-sentential expressions are determined by the contributions these expressions make to the inferential roles the sentences containing them participate in (see Dummett, 1991; Brandom, 1994).

The inferentialist mode of semantic explanation might be thought to apply *locally*, that is, to restricted regions of language such as logical, moral, causal, deontic, epistemic, or theoretical terms; or it might be advanced as a *global* thesis according to which the inferential model of explanation extends to language at large. In order to be able to accommodate expressions that are less straightforwardly explained in inferential terms, global inferentialists will have to include 'language-entry' and 'language-exit' rules (Sellars, 1953) among the relevant 'inferential' connections. That is, uses of expressions like observational predicates will have to be linked to perceptual cues (on the 'entry' side), and they must link up with intentional action (on the 'exit' side). In the absence of any such anchoring of meaning in our experience and interaction with the world, our inferential language game threatens to fail to "latch onto the world"; it could not serve as a means for representing the world. We will be examining local versions of IRS in §2, while global accounts of IRS will occupy us in §3.

The comparison of IRS with informational semantics is telling also in other respects. For one, it may be noted that informational semantics naturally entrains semantic atomism. As Fodor (1990) has stressed, it is possible (at least in principle) on such views that a creature should possess but one concept. Since concept possession depends on being suitably related to the environment, having a concept need not presuppose the possession of others. Things look different when viewed through the prism of IRS, which is incompatible with semantic atomism. The content of a concept is determined by its connections to other suitably inferentially related concepts. Hence, one cannot so much as have a concept without having many concepts.[5] IRS is thus compatible with, but does not mandate, semantic holism – the position that a statement's meaning (and derivatively the meanings of the sub-sentential expressions figuring within it) is determined by the entire network of inferential connections in which it participates. Whether a particular form of IRS endorses holism and how far-reaching that holism is will depend on which types of inferences are taken to be determinative of meaning. If all inferential links – mediate as well as immediate ones – are to taken into account, IRS will amount to an all-out or 'pandemic' holism: the meaning of one statement is fixed by its place in the network of inferential connections linking it to all other statements expressible in the language. By contrast, some versions of IRS treat only certain more restricted subclasses of inferences to be determinative of meaning: analytic inferences or inferences that display a particular counterfactual robustness (Sellars, 1953). Michael Dummett's 'molecularism' is, as the name makes plain, a further attempt at carving out a middle ground between atomism and holism. The molecularist allows for what we might call *semantic clusters*. A semantic cluster is a set of expressions or phrases that are mutually dependent in the sense that the meaning of any member of the set is determined by its inferential links to all the others in the set. Examples are groups of contrary predicates like color words or phrases like 'mother of,' 'father of,' 'child of.' Famously, according to Quine the expressions 'analytic,' 'necessary,' 'synonymous' also form a cluster. What the molecularist opposes is the holist notion that all of these clusters collapse into one all-encompassing master cluster, language as a whole.

Typically, strongly holistic versions of IRS are less apt to answer the epistemological question underlying UND: 'What is it that a speaker must know in order to count as understanding an expression?' For, while all-out holism may be a candidate explanation of how it is, metaphysically speaking, that expressions of a language have the meanings they do, it does not deliver a plausible criterion of linguistic understanding. Transferred to an epistemological key, a strongly holistic version of IRS amounts to the thesis that a speaker must somehow grasp the entirety of the inferential network of links obtaining between all the statements expressible in the language. A claim that seems incredible. Not only would no human speaker of any non-trivial language qualify as understanding any expressions of her language, it also seems deeply unintuitive that my understanding of 'measurable cardinal' (say) should be tied to my appreciation of the correctness of the inference from the proposition that Puck is a cat to the proposition that Puck is an animal. Thus, advocates of IRS *qua* account of linguistic understanding are likely to opt for a non-holistic variant of IRS.[6] The same goes for philosophers who, like Dummett, believe that MD and UND are inseparable, that a theory of meaning is of necessity a theory of what it is a speaker must know in order to understand the expressions of the language. In the following, when we wish to speak of the meaning-determinative or understanding-constituting class of inferential connections in a way that is neutral with respect to the degree of holistic dependence, we speak of the *salient class* of inferences.

Another important choice point is whether IRS is to be interpreted *descriptively* or *normatively*. On a descriptive reading of IRS, the salient class of inferences is constituted by the inferential connections speakers actually make or are disposed to make or accept under various actual and counterfactual circumstances.[7] On a normative reading, MD and UND are understood in terms of the inferential connections we *ought* to be disposed to make or accept. As Daniel Whiting (2006) emphasizes, the normativity in question is not merely consequent upon an expression's having a certain meaning. For instance, that 'cat' means *cat* may have the implication that one ought to apply 'cat' to all and only cats. The case of normative versions of IRS is different. The normativity of the salient class of inferential roles is part and parcel of the meanings they determine: it is because one is in some appropriate sense disposed to recognize the propriety of the inferential connections that one (the community) means what one (it) does by a given expression. Similarly, understanding an expression consists in recognizing the propriety of the relevant set of inferential connections.[8]

This brings us back to another important distinction: the distinction, namely, between individualistic and social (or anti-individualistic) interpretations of IRS. MD, for instance, might be thought of as a thesis about the determination of meanings of a particular speaker's idiolect (at a particular time), or it might be interpreted as a thesis about how the meanings of a public language are fixed. The former thesis is compatible with internalistic conceptions of meaning according to which the determinants of meaning are intrinsic properties of the speaker. Aside from all-out internalism, such views are compatible also with so-called two-factor views. Two-factor views distinguish two aspects of meanings: narrow and broad content. While broad content may be partly determined by the speaker's social or physical environment, narrow content is understood internalistically. Within such a two-factor view, the individualistic reading of an IRS-based thesis about meaning determination might be appealed to as an account of narrow content. In contrast, the anti-individualistic interpretation of IRS is compatible with a social externalism (Burge,

1979; 1986; Putnam, 1981). It is less clear that such views can accommodate twin-earth-type arguments in favor of physical externalism – the view that the meanings of certain expressions, typically proper names and natural kind terms, are in part individuated by their physical environment. For example, in Putnam's classic thought experiment (Putnam, 1975) the meaning-determining inferential roles for 'water' are identical in Oscar's and in Twin Oscar's linguistic communities. All the same, 'water' presumably picks out different substances and hence has different meanings (assuming that meanings determine referents). Some versions of *conceptual* role semantics (incorporated into one-factor views of content) allow for so-called 'long-arm roles,' which 'reach out,' so to speak, to include the speaker's physical environment and her causal interactions within the speaker's conceptual roles (Harman, 1987). Some accept such 'long-arm roles' within IRS. Advocates of IRS who do not – while maintaining that inferential roles determine reference and truth-conditions – will either have to restrict IRS to types of expressions where the physical externalist intuitions have less of a foothold or they will have to dispute those intuitions altogether.

Now, we have said that some proponents of IRS are motivated by MD. But what does it mean to say that inferential roles determine meanings? A fully worked out account of IRS must clarify two things: the relation of determination involved, and the notion of meaning appealed to. Begin with the first of these tasks. On perhaps the most straightforward reading, meanings might simply be identified with inferential roles or rules of inference. Alternatively, meanings might be taken to supervene on inferential roles in the sense that identity of inferential roles guarantees identity of meanings. Or perhaps there are other more sophisticated ways in which inferential roles determine meanings (we will consider one such account in §3).

The second task is not independent of the first. Whether meanings can be identified with inferential roles, for instance, will depend on what meanings are. We consider four uses of 'meaning.' First, 'meaning' may be used to designate the referent or extension of an expression. Second, the meaning of an expression may be equated with the compositional contribution it makes to what is said by a sentence in which it is being used. Third, meanings have sometimes been thought to be the semantic determinants of the referent or extension of an expression. On simple descriptivist accounts, the meaning of a proper name (say) may be a definite description (or cluster of such descriptions). It is in virtue of the description's being satisfied by a particular object that the associated name names its bearer. Fourth, and finally, 'meaning' is often used to designate what it is that a speaker must grasp in order to understand it.

Here is not the place for a comprehensive survey of all of the combinatorial possibilities. Nevertheless, a handful of examples will give the reader an impression of the clarificatory work necessary to fully spell out a version of inferentialism. For instance, inferential roles cannot be identified with meaning in the first sense of 'meaning,' since what we talk about when we talk about Saul Kripke or aardvarks are people and perhaps kinds, properties, or classes of objects, not inferential roles. Similarly, if meaning is understood as the semantic contribution to what is said by a sentence, then if what is said by a sentence is a Russellian proposition or a proposition conceived of as a set of possible worlds, inferential roles, again, cannot be identified with meanings. It is at least conceivable, by contrast, that inferential roles are (or otherwise determine) that which determines reference, or that they are what a speaker needs to master in order to understand an expression having those inferential roles.

Completing the survey would require a similar analysis of other forms of meaning determination.

Enough, then, about IRS's place in the landscape of contemporary approaches in the philosophy of language. The next two sections focus on, respectively, arguably the most important local version of IRS, *logical* inferentialism, and Brandom's global inferentialism.

2 Logical Inferentialism

Though various local brands of IRS have been advanced, logical inferentialism, IRS as it applies to logical vocabulary, deserves special attention. Inferentialist accounts of logical expressions, we noted, seem especially natural. As a result, such accounts are often regarded as a model for IRS in general. As Brandom puts it, typically inferentialists "look to the contents of logical concepts as providing the key to understanding conceptual content generally" (Brandom, 2007, p. 653). We begin by introducing some of logical inferentialism's standard motivations and commitments (§§2.1–2.2). We then discuss a response inspired by Gerhard Gentzen's work to Arthur Prior's famous attempt to undermine the view, and consider a potential corollary of Gentzen's response, *viz.* that logical inferentialism validates a non-classical logic (§§ 2.3–2.4).

2.1 The Only Game in Town?

Inferentialists frequently distinguish two central aspects of the correct use of a sentence: the conditions under which it may be correctly asserted, and the consequences that may be correctly derived from (an assertion of) it. As Dummett puts it:

> crudely expressed, there are always two aspects of the use of a given form of sentence: the conditions under which an utterance of that sentence is appropriate, which include, in the case of an assertoric sentence, what counts as an acceptable ground for asserting it; and the consequences of an utterance of it, which comprise both what the speaker commits himself to by the utterance and the appropriate response on the part of the hearer, including, in the case of assertion, what he is entitled to infer from it if he accepts it. (Dummett, 1973, p. 396)

On their most common interpretation, introduction rules in a natural deduction system (henceforth, I-rules) state the sufficient, and perhaps necessary, conditions for introducing logically complex sentences of the corresponding kind; elimination rules (henceforth, E-rules) tell us what can be legitimately deduced from any such sentence. Logical inferentialism, then, becomes the claim that the meanings of logical expressions are fully determined by their I- and E-rules (corresponding to MD above), and that to understand such expressions is to use them according to such rules (corresponding to UND).[9]

The idea that rules can fix meanings became increasingly popular in the 1930s and 1940s. For logical expressions, the strategy is an especially tempting one. For one thing, one can *prove* that, if $\wedge$ satisfies its I- and E-rules

$$\wedge\text{-}I\frac{A \quad B}{A \wedge B} \qquad \wedge\text{-}E\frac{A \wedge B}{A} \quad \frac{A \wedge B}{B}$$

and if the rules are truth-preserving, then sentences of the form $\ulcorner A \wedge B \urcorner$ must have their standard truth-conditions:

$$(\wedge)\ \ulcorner A \wedge B \urcorner \text{ is true iff } \ulcorner A \urcorner \text{ is true and } \ulcorner B \urcorner \text{ is true.}^{10}$$

For another, a speaker who did not master $\wedge$-I and $\wedge$-E can hardly be credited with an understanding of conjunction. And, conversely, it would seem to be a mistake not to attribute an understanding of conjunction to a speaker who *did* master $\wedge$-I and $\wedge$-E. Indeed, what else could account for one's understanding of logical expressions? As Paul Boghossian puts it:

> It's hard to see what else could constitute meaning conjunction by 'and' except being prepared to use it according to some rules and not others (most plausibly, the standard introduction and elimination rules for 'and'). Accounts that might be thought to have a chance of success with other words – information-theoretic accounts, for example, or explicit definitions, or teleological accounts – don't seem to have any purchase in the case of the logical constants. (Boghossian, 2011, p. 493)

Accordingly, the view that for logical expressions inferentialism is the only game in town, is widely shared.

Boghossian offers a second argument in favor of logical inferentialism, *viz*., that it makes for an elegant account of blind but blameless reasoning – one that seeks to explain how justification (or knowledge) can be transmitted from premises to the conclusion in deductive inference. In a nutshell, any such account is constrained by the failures of simple inferential internalism (SII) and simple inferential externalism (SIE). According to SII, it is required not only that (i) I be justified in believing the premises of a deductive inference and (ii) that the conclusion be justified independently of the premises, but also (iii) that I can know by reflection alone that the premises provide me with good grounds for believing the conclusion. SII amounts to a form of access internalism about deductive inferential justification. Aside from (i) and (ii), SIE requires (iv) that the pattern of inference exemplified be valid (necessarily truth-preserving). Boghossian's proposal is to suggest instead that our inferences are *blind*, because we cannot be expected to satisfy (iii), on pain of starting an infinite regress or else invoking dubious faculties of rational insight. But it is also not enough that our inferences satisfy (iv) since inferences may be reckless and hence blameworthy despite being truth-preserving. So, justification transferral must admit of blind and blameless inferences:

> a deductive pattern of inference *P* may be blamelessly employed, without any reflective appreciation of its epistemic status, just in case inferring according to *P* is a precondition for having one of the concepts ingredient in it. (Boghossian, 2003a, p. 239)

The fact that I take *A* thoughts to be a warrant for believing *B* thoughts is constitutive of my having these thoughts (*A* or *B*) at all. But, then, how can I be epistemically blameworthy for making such an inference? We return to this admittedly controversial argument in §4 below.

The resulting view, then, is an analytic approach to logic – one according to which logical truths are *epistemically analytic* (Boghossian, 1996; 2003b): if *A* expresses a logical truth, then the proposition it expresses can be known on the basis of a grasp of the meaning of the

sentence alone. Whether logical inferentialism is also committed to a *metaphysical* conception of analyticity – one according to which logical truths owe their truth solely to the meanings of the logical expressions (and to the facts) – is more controversial (see, e.g., Russell, 2014; Warren, 2015).

2.2 Harmony

Logical inferentialism is a form of *conventionalism*: certain logical laws, the thought goes, are stipulated to hold or else are somehow extracted from our practice. Thus, Dummett writes:

> Although it is not true of logical laws generally that we are entitled simply to stipulate that they shall be treated as valid, there must be certain laws or systems of laws of which this holds good. Such laws will be 'self-justifying': we are entitled simply to stipulate that they shall be regarded as holding, because by so doing we fix, wholly or partly, the meanings of the logical constants that they govern. (Dummett, 1991, p. 246)

The thought seems to be this: while in the overwhelming majority of cases the question whether we may accept a certain logical law is already settled (it depends on whether the given law can be derived from laws that are already accepted), *basic* laws cannot be justified in this way, on pain of an infinite regress. The question arises, however, whether *any* seemingly basic law can be regarded as determinative of meaning.

An affirmative answer yields disaster, as Arthur Prior's infamous binary connective `tonk` shows (see Prior, 1960):

$$\texttt{tonk-I}\frac{A}{A\ \texttt{tonk}\ B} \qquad \texttt{tonk-E}\frac{A\ \texttt{tonk}\ B}{B}.$$

If the consequence relation is transitive, and at least one theorem can be proved in one's system, then *any* sentence can be proved in one's system. To the best of our knowledge, the first sketch of an inferentialist solution to the problem was given by Gerhard Gentzen in 1934. In a famous passage, Gentzen writes:

> To every logical symbol &, ∨, ∀, ∃, →, ¬, belongs precisely one inference figure which 'introduces' the symbol – as the terminal symbol of a formula – and which 'eliminates' it. The fact that the inference figures &–E and ∨–I each have two forms constitutes a trivial, purely external deviation and is of no interest. The introductions represent, as it were, the 'definitions' of the symbols concerned, and the eliminations are no more, in the final analysis, than the consequences of these definitions. This fact may be expressed as follows: in eliminating a symbol, we may use the formula with whose terminal symbol we are dealing only 'in the sense afforded it by the introduction of that symbol'. (Gentzen, 1934, p. 80)

Gentzen argues that the I-rules of his newly invented calculus of natural deduction 'fix,' or 'define,' the meanings of the expressions they introduce. He also observes that, on this assumption, E-rules cannot be chosen randomly. They must be justified by the corresponding I-rules: they are, in some sense, their 'consequences.' This key thought expresses *in nuce* the idea that I- and E-rules must be, in Dummett's phrase, in *harmony* with each other.

Conversely, if it is thought that E-rules are meaning-constitutive, I-rules cannot be chosen arbitrarily either (see, e.g., Dummett, 1991, p. 215).

This intuitive idea can be spelled out in a number of ways. Dummett (1991, p. 250) and Prawitz (1974, p. 76) define harmony as the possibility of eliminating *maximum formulae* or *local peaks*, that is, formulae that occur both as the conclusion of an I-rule and as the major premise of the corresponding E-rule (see also Prawitz, 1965, p. 34). The following reduction procedure for $\rightarrow$, for instance, shows that any proof of B via $\rightarrow$-I and $\rightarrow$-E can be converted into a proof from the same or fewer assumptions that avoids the unnecessary detour through (the introduction and elimination of) $A \rightarrow B$:

$$\dfrac{\dfrac{\begin{matrix}\Gamma_0,\,[A]^i\\ \Pi_0\\ B\end{matrix}}{A \rightarrow B}\,{\scriptstyle \rightarrow\text{-I},\,i} \qquad \begin{matrix}\Gamma_1\\ \Pi_1\\ A\end{matrix}}{B}\,{\scriptstyle \rightarrow\text{-E}} \quad \rightsquigarrow_r \quad \begin{matrix}\Gamma_1\\ \Pi_1\\ \underbrace{\Gamma_0, A}\\ \Pi_0\\ B\end{matrix}$$

where $\rightsquigarrow_r$ reads 'reduces to.' Dummett (1991, p. 250) calls the availability of such procedures *intrinsic* harmony where, crucially, the reduction reduces the *degree of complexity* of a derivation, that is, the number of occurrences of logical operators. He correctly points out, though, that it only prevents elimination rules from being stronger than the corresponding introductions, as in the case of Prior's `tonk`. It does not rule out the possibility that they be, so to speak, too weak (see Dummett, 1991, p. 287).[11] A way to ensure that E-rules be strong enough is to require that they allow us to *reintroduce* complex sentences, as shown by the following *expansion*:

$$\begin{matrix}\Pi\\ A \wedge B\end{matrix} \quad \rightsquigarrow_e \quad \dfrac{\dfrac{\begin{matrix}\Pi\\ A \wedge B\end{matrix}}{A}\,{\scriptstyle \wedge\text{-E}} \qquad \dfrac{\begin{matrix}\Pi\\ A \wedge B\end{matrix}}{B}\,{\scriptstyle \wedge\text{-E}}}{A \wedge B}\,{\scriptstyle \wedge\text{-I}}$$

where $\rightsquigarrow_e$ reads 'can be expanded into.' This shows that any derivation Π of $A \wedge B$ can be expanded into a longer derivation which makes full use of both $\wedge$-I and $\wedge$-E. Accordingly, a pair of I- and E-rules for a constant \$ can be taken to be harmonious iff there exist both reduction and expansion procedures for \$-I and \$-E. Alternative conceptions of harmony are developed in, for example, Read (2000) and Tennant (1997; 2008). For an overview see Steinberger (2011a). But *why* exactly should I- and E-rules for logical expressions be harmonious?

One motivating thought behind the requirement of harmony is that logic is *innocent*: it shouldn't allow us to prove atomic sentences that we couldn't otherwise prove by first introducing and subsequently eliminating a logic operator (Steinberger, 2009). Yet another motivating thought has it that I-rules determine, in principle, necessary and sufficient conditions for introducing complex sentences. The necessity part of this claim is in effect what Dummett calls the Fundamental Assumption, that "[i]f a statement whose principal operator is one of the logical constants in question can be established at all, it can be established by an argument ending with one of the stipulated I-rules" (Dummett, 1991, p. 252). The Assumption lies at the heart of proof-theoretic accounts of validity (Prawitz, 1985; Dummett, 1991). To see that it justifies a requirement of harmony, let $CG[A]$ be the canonical grounds for a

complex statement A, as specified by its I-rules. Then, we may reason that, by the Fundamental Assumption, B follows from $CG[A]$ if and only if B follows from A itself.[12] In short: it is a consequence of the Fundamental Assumption that complex statements and their grounds, as specified by their I-rules, must have the same set of consequences. That is, I- and E-rules must be in harmony with each other in the following sense: one may infer from a complex statement nothing more, and nothing less, than that which follows from its I-rules.

If harmony is a necessary condition for logicality, then Prior's `tonk` need not worry the logical inferentialist: the `tonk` rules are spectacularly disharmonious, and hence cannot define a *logical* connective.[13] The rules are also *non-conservative*: they allow one to prove sentences in the `tonk`-free language that were not previously provable in the absence of the rule for `tonk` (they allow us to prove any such sentence). And indeed, the first response to Prior's `tonk`, published by Nuel Belnap in 1962, was precisely that admissible rules should yield conservative extensions of the base systems to which they may be added.[14]

The conservativeness requirement is equivalent to the requirement that an admissible logical system S be *separable*, that is, such that every provable sentence or rule in the system has a proof that only involves either structural rules or rules for the logical operators that figure in that sentence or rule. In conjunction with UND, separability makes for an *atomistic* account of one's understanding of the logical vocabulary – one according to which to understand \$ is to know how to use it correctly; the totality of uses of \$ (i.e., the derivations of rules and theorems involving sentences with \$ as their main logical operator) is derivable from the basic rules for \$, and, consequently, one's grasp of \$'s rules is sufficient for knowing \$'s meaning. Thus, on the foregoing view, a speaker could understand $\wedge$ without understanding $\exists$, $\rightarrow$ without understanding $\neg$, and so forth. In so far as our understanding of the logical vocabulary *could* be atomistic, it might be argued that an adequate axiomatization of logic ought to be separable, on pain of ruling out a possible way our understanding actually is.

2.3 *Inferentialism and Logical Revision*

Proof-theoretic constraints such as harmony, conservativeness, and separability rule out Prior's `tonk`. However, it may be argued that they rule out much more. For while the rules of intuitionistic logic are harmonious, standard formalizations of classical logic typically aren't (Dummett, 1991; Prawitz, 1977; Tennant, 1997). For instance, the classical rule of double negation elimination

$$\text{DN}\,\frac{\neg\neg A}{A}$$

is not in harmony with the standard rule of negation introduction:

$$\neg\text{-I}, i\,\frac{\begin{array}{c}[A]^i\\ \vdots\\ \bot\end{array}}{\neg A}.$$

The harmonious rule of negation elimination is the following *intuitionistic* rule:

$$\neg\text{-E}\,\frac{A \quad \neg A}{\bot}.$$

This rule, unlike its classical counterpart, allows us to infer from $\neg A$ precisely what was required to assert $\neg A$: a derivation of $\bot$ from A. But, then, double negation elimination is left, so to speak, in the cold. Similarly, standard axiomatizations of classical logic are not separable. For instance, some uses of $\rightarrow$ such as Peirce's Law, that $((A \rightarrow B) \rightarrow A) \rightarrow A$, are only derivable by means of rules for *both* $\rightarrow$ and $\neg$. Intuitionists such as Dummett, Prawitz, and Tennant have taken this observation to show that classical rules such as double negation elimination are not logical (or that they are in some other sense defective), and that the logical rules we should adopt are those of *intuitionistic logic*, that is, classical logic without the Law of Excluded Middle ($A \vee \neg A$), double negation elimination, and other equivalent rules (or perhaps of a weaker logic still; Tennant, 1987; 1997).

This argument is problematic, however. For while it is true that *standard* axiomatizations of classical logic are not harmonious, a number of non-standard axiomatizations are both harmonious and separable. In particular, classical logic can be shown to be as proof-theoretically as respectable as intuitionistic logic provided rules are given both for asserting and for *denying* complex statements (Rumfitt, 2000; Incurvati and Smith, 2010), where denial is taken to be a primitive speech act distinct from the assertion of a negated sentence (Parsons, 1984; Smiley, 1996). The negation rules for classical negation are then as harmonious as the intuitionistic ones: they allow one to deny $\neg A$ given the assertion of A and *vice versa*, and to deny A given the assertion of $\neg A$ and *vice versa*.[15]

Local forms of inferentialism, beyond logical inferentialism, have recently been developed by a number of authors. For instance, Matthew Chrisman (2010; 2015) and Ralph Wedgwood (2007) develop an inferentialist account of deontic modals, Julian Reiss (2012) offers an inferentialist account of causal claims, Mauricio Suarez (2004) defends an inferential conception of scientific representation. In the next section, we consider the most prominent attempt at elaborating a global inferentialism for language at large.

3 Brandom's Inferentialism

Perhaps the most worked out version of global inferentialism has been put forward by Robert Brandom (1994; 2000; 2008).[16] In this section, we summarize Brandom's account and situate it with respect to the choice points within inferentialism identified in §2. Given the expanse of Brandom's writings and the critical literature it gave rise to, our discussion will inevitably have to omit many noteworthy aspects of Brandom's work and of the responses to it.

Brandom's account, like any inferentialist account, is use-theoretic or, in his preferred vocabulary, pragmatic. Semantics, in Brandom's slogan, is 'answerable to pragmatics' (Brandom, 1994, p. 83). His account's point of departure is the *doings* of linguistically endowed creatures, in particular their practices of asserting and inferring which, according to him, 'come as a package' (Weiss and Wanderer, 2010 p. 313). It is through the speech act of assertion that we advance claims which, in turn, are expressed by declarative sentences. And it is the contents of these assertions (expressed by declarative sentences) that are susceptible of standing in inferential relations. It is for this reason that declarative sentences are taken to be the primary unit of significance and so enjoy a privileged semantic standing (Brandom, 1994, p. 79).

Aside from being a use theory of meaning, Brandom's theory may further be classified as an assertibility theory of meaning: the meanings of declarative sentences are to be

explained in terms of the conditions under which an assertion is appropriate or correct. However, Brandom's approach differs in a number of significant ways from other types of assertibility theories.

Brandom follows Dummett's lead in seeking to extend to the whole of language Gentzen's bipartite model of the meanings of logical expressions in terms of I-rules and E-rules (see §2). On the proposed picture, the meaning of a sentence (or of a thought) can be characterized in terms of the two aspects of its assertoric use: the 'set of sufficient conditions' that would warrant its assertion and the 'set of necessary consequences' (Brandom, 2000, p. 63) of doing so (as well as, we will see, the set of claims incompatible with it).

Crucially, the inputs for and the outputs of assertions are construed inferentially. That is, the meaning of a sentence is to be explained via its inferential antecedents and consequences. The meanings of sub-sentential expressions, though unfit in and of themselves to act as relata of inferential connections, are then to be accounted for in terms of the systematic contributions they make to the assertibility conditions of the sentences of which they are constituents. Their semantic contributions are accounted for in terms of their substitutional behavior (Brandom, 1994, ch. 6). Where Brandom departs from Dummett is in the deflationary nature of his account. Traditional semantic notions of reference and truth are explained in broadly deflationist terms, as having an 'expressive' role – they enable us to render explicit certain intentional and anaphoric features of our practice which otherwise remain implicit within them (Brandom, 1994, ch. 5).[17]

As we noted in §2, certain expressions appear to lend themselves more readily to inferentialist treatment than others. However, since Brandom's account lays claim to global applicability, it must be capable also to account for observational predicates like 'red,' for instance. According to Brandom, even the possession of a color concept like *red* requires more than merely an ability to respond differentially to red things. Parrots, barometers, and thermometers all respond differentially to certain stimuli, yet they cannot be credited with the corresponding concepts. Full conceptual competence presupposes an appreciation of the inferential connections from and to thoughts containing the concept in question, for example, the inference from 'this is crimson' to 'this is red,' or from 'this is red' to 'this is colored.'

What confers meaning on a sentence is that it may be correctly inferred from certain sentences, and that other sentences may be inferred from it. This presupposes an appreciation on the part of the speaker that the sentence may act as a premise and as a conclusion in arguments.[18] In Brandom's oft-cited Sellarsian slogan, the meaning-determinative linguistic and conceptual practices of asserting and inferring are to be conceived of as taking place within the 'game of giving and asking for reasons.' His aim is to explain how our practices of asserting, challenging, defending, and retracting assertions by exploiting the inferential connections within which the asserted contents stand, are able to confer meanings on the linguistic vehicles by which these linguistic acts are performed.

Crucially, the meaning-constitutive inferential patterns are not *formal* inferences (i.e., deductive relations that are truth-preserving in virtue of the logical forms of the claims), but rather *material* ones including analytic inferences like 'Philadelphia is south of New York City; therefore, New York City is north of Philadelphia' as well as defeasible *a posteriori* ones like 'this substance burns in a white flame; so, this substance is magnesium.' It is these material inferential relations (not formal ones) that we rely upon in justifying, challenging, and defending our assertions, and it is therefore they which constitute inferential roles of sentences.[19] The supposed primacy of formal inference is taken to be an artifact of the

referentialist order of explanation: sub-sentential expressions are thought to have semantic values in virtue of the referential relations they stand in. These semantic values compositionally determine the truth-values of sentences. And these, together with the logical forms of the sentences, go on to determine the formal logical consequences of those sentences. On Brandom's account, by contrast, it is material inference that takes center stage. Indeed, logical concepts are inessential to (non-logical) conceptual practices. The role of logical vocabulary, as that of semantic vocabulary, is *expressive*, according to Brandom. It is in virtue of such expressive vocabulary that we are able to subject our concepts to critical scrutiny, laying bare the commitments we incur by virtue of operating with those concepts. However, it is not directly determinative of meaning.[20]

An important feature of Brandom's inferentialism, we said, was the bipartite structure of its analysis of assertibility conditions, which it inherits from Dummett. Like Dummett, Brandom rejects assertibility theories that focus on the grounds of assertion to the exclusion of the consequences of assertion. Such one-sided accounts, he claims, are unable to discriminate the meanings of 'I will write a book about Hegel' and 'I foresee that I will write a book about Hegel.' For while the circumstances warranting the assertion of either sentence are the same, the consequences of doing so differ Brandom (2000, p. 65). One-sided accounts, according to him, are unable to account for that difference.

Also, much like Dummett, Brandom's inferentialism is a broadly social and normative inferentialism. The relevant meaning-determining inferential roles are the ones that govern not the expressions of a particular idiolect but those of a public language. What is more, inferential roles, for Brandom, are not to be construed merely as *regularities* within a social inferential practice, but rather as *proprieties* and so normatively. Where Brandom's assertibility theory goes beyond Dummett's is in that he aims to offer an analysis of the 'normative fine structure' of the grounds and the consequences of asserting in terms of the normative categories of *commitment* and of *entitlement*.

Start with commitment. Asserting is a way of expressing one's endorsement of a proposition. But for such an endorsement to have the distinctive force of an assertion, it must be subject to the norm that assertions entail commitments to the material implications of the asserted contents. A speech act not accompanied by such a commitment would not qualify as an assertion. For instance, in asserting that this is red, I commit myself to the claim that this is colored. Commitment thus has the deontic force of obligation: having asserted that this is red, I ought also to endorse what follows from the content of my assertion, for example, that this is colored.

The second category of entitlement must be understood against the backdrop of Brandom's rationalism about language. The practice of asserting, for him, is essentially bound up with our practice of exchanging reasons. Upon advancing a claim by asserting it, I may appropriately be challenged by you. I meet your challenge by demonstrating my entitlement to the claim, for instance by pointing to a further claim to which I am already entitled and from which the claim I advance may be correctly inferred. For example, I can demonstrate my entitlement to the claim that this is red provided that I am entitled to the claim that this is scarlet. The reason my entitlement carries over from an assertion of the latter sentence to an assertion of the former, is because the sentences express contents that stand in an entitlement-preserving inferential relation. While commitments corresponded to the 'deontic status' of obligations, we can now see that entitlements correspond to permissions.

The deontic modalities (like other modalities) of obligation and permission are customarily taken to be duals of one another as witnessed by their interdefinability with the help of

formal negation (e.g., it is obligatory that *p* just in case not-*p* is not permissible and vice versa). Given his conception of logical vocabulary as having an expressive and hence auxiliary role, Brandom does not wish to appeal to the formal concept of negation at this explanatorily fundamental level. Nevertheless, commitments and entitlements interact in important ways. In particular, Brandom defines a notion of incompatibility or material negation in terms of them: two propositions are incompatible with one another just in case commitment to one precludes entitlement to the other.

Aside from being able to give a more refined account of the central notion of inferential role, Brandom claims that the three additions to his conceptual tool belt – commitment, entitlement, and incompatibility – enable him to deal with a problem that has long bedeviled assertibility theories of meaning. To see what the trouble is note that assertions can be said to be correct in two ways. An assertion can be said to be correct *by the lights* of the agent or of the agent's linguistic community if it is warranted by the relevant standards (e.g., 'Was the evidence properly taken into account?' 'Were the inferences made sound ones?' and so on). But of course even though the individual may be subjectively correct and hence blameless, he may still be wrong. So there is a second, objective sense in which an assertion's correctness may be appraised. In this sense an assertion is correct just in case it is true, that is, just in case it tells it how it is. Traditionally, Brandom claims, assertibility theories have struggled to account for objective correctness in this sense.

Oftentimes the assertibility theorists resorted to certain types of idealizations in order to narrow the gap between the two types of normative appraisals by defining objective correctness as subjective correctness under certain ideal conditions (perfect evidence, at the end of inquiry, etc.). Brandom believes that idealizing maneuvres of this kind are doomed to failure. For a typical problem case consider Brandom's example:

(1) The swatch is red.
(2) The claim that the swatch is red is now assertible by me.

Although the second sentence merely seems to make explicit what according to the assertibility theorist is implicit in the act of asserting the first sentence, the two sentences intuitively differ in content. The two sentences, though assertible in the same circumstances, differ in their truth-conditions. However, Brandom believes that his account delivers the means necessary to capture the difference in semantic content without abandoning the assertability. For the two sentences to have the same content, they would have to rule out the same claims; they would have to be, in Brandom's vocabulary, 'incompatibility-equivalent' (Brandom, 2000, p. 199). But this is not the case. For instance, (1) is compatible with it being the case that rational beings have never evolved, whereas (2) is not.[21]

Needless to say, Brandom's grand project has faced a great number of criticisms.[22] Here we will single out merely one central strand of criticisms because of its relevance to other inferentialist enterprises: Brandom's holism.[23] In §2, we noted that if inferentialism is to play the part of a meta-semantic account – a theory of MD, of UND, or both – it must explain which types of inferences have the relevant meta-semantic relevance; that is, which inferences are meaning-determinative or which ones are understanding-constitutive. Brandom, we have said, endorses the controversial thesis that *all* inferential connections to which the sentence contributes are relevant. We have already noted some of the intuitive difficulties faced by Brandom's view *qua* theory of understanding. In the following passage Brandom

himself stresses in particular the difficulty of accounting, on a holistic account, for the 'possibility of communication or of interpersonal understanding':

> If the inferential significance of a claim depends on what else one is committed to, then any difference between the collateral commitments of speaker and audience can mean that a remark has a different significance in the one's mouth than it does in the other's ear. How is it then possible to make sense of the idea that they understand one another, so as to be able to agree or disagree? If the contents expressed by sentences must be individuated as finely as the theories they are embedded in, the intelligibility of communication across theories – the very notion of conveying information – is threatened. […] If, because of his very different collateral commitments, Rutherford meant something quite different by 'electron' than I do, it seems I can't disagree with him about whether electrons have fixed positions and orbits, since I can't either say or think anything with the content he would have expressed by saying "electrons orbit the nucleus." How, then, are we to understand so much as the possibility of cognitive progress in science? (Brandom, 2007, p. 666)

How, in the light of this is a holistic account of understanding and communication to get off the ground? A natural answer, it might be thought, is to maintain that Rutherford's and Brandom's use of 'electron' overlap sufficiently to ensure communication; that their uses of the term are sufficiently similar. However, Fodor and Lepore (2001) and others have argued that such appeals to similarity will not do. The reason, according to them, is that any appeal to similarity in use would appeal to related expressions. For instance, it is of no help to point to the fact that both Brandom and Rutherford may agree that electrons are negatively charged, because their different theoretical commitments lead them to assign different meanings to 'charge' (see Fodor and Lepore, 2001). However, Brandom takes this conclusion to be too hasty. While it may be that many of the inferential roles the two associate with the expression differ in myriad ways, many of the non-inferential circumstances under which both would apply the term are the same, or at least sufficiently similar:

> Thus Rutherford and I are both disposed to respond to a bolt of lightning by applying the term 'electron', and to respond to applying the expression 'high voltage, high amperage electron flow' to a bare piece of metal by avoiding contact with it. These language entry and language exit moves, no less than the language-language ones, also give us something important in common, even when described at a so-far-subsemantic level, that is, in a nonsemantic vocabulary. I do not see why the structures so-described do not underwrite a perfectly intelligible notion of partially shared, or merely similar inferential roles. (Brandom, 2007, p. 666)[24]

Moreover, Brandom dismisses the notion that understanding should be understood on the 'Lockean' model according to which my understanding you consists in your idea being transferred or reproduced in my mind. Instead, we should understand 'understanding' in accordance with Brandom's picture of the normative practices that are constitutive of meanings in the first place. One's advancing a sentence assertorically should be understood against the background of social 'scorekeeping' practices in which my interlocutors track my and others' commitments and entitlements as a result of my assertion (see Brandom, 1994, pp. 180ff.). These scorekeeping practices represent Brandom's account of how it is that content-conferring norms are socially instituted. They are thus to be understood as an

attempt at a reductive (though non-naturalistic) account of content. Contents (and intentionality with them) are to be understood in terms of the normative practices by which we monitor and assess our rational discourses of asserting, challenging, justifying, and retracting. "The capacity to understand each other," on this picture, "is the practical ability to navigate across the gulf between doxastic perspectives created by the effect of differing collateral commitments on the inferential significance of one noise in the mouth that utters it and the ear that hears it" (Brandom, 1994, p. 667). The trouble with this reductive view, however, is that it seems hard to see how explanatorily basic normative practices can be described without already making use of intentional vocabulary (Fodor and Lepore, 2010; Rosen, 1997).

Rather than further dwelling on objections specifically leveled at Brandom, we now turn to a number of more general objections to inferentialism.

4 Objections and Replies

Objections to IRS are legion.[25] For reasons of space, we focus on two main lines of criticism: a general worry about IRS *qua* unorthodox semantics, and Timothy Williamson's recent sustained attack to UND, understood as the claim that to understand an expression is to be disposed to make and accept basic inferences featuring it. We begin with the former.

The thought that the content of a sentence in context is given by, perhaps among other things, its truth-conditions – truth-conditional semantics, for short – lies at the heart of much of contemporary linguistic semantics and philosophy of language. As Williamson observes, "this simple idea has been basic to the massive development of mainstream formal semantics over recent decades, in both linguistics and philosophy of language, for natural and artificial languages" (Williamson, 2010).[26] However, Williamson argues, this simple and fruitful idea is incompatible with IRS. He writes:

> If you want an explicit theory of how some particular linguistic construction contributes to the meanings of sentences in which it occurs, the inferentialist is unlikely to have one. Better try the referentialist. (Williamson, 2010, p. 23)[27]

The thought seems to be this. Given the empirical success of contemporary truth-conditional semantics, it would be a mistake to abandon it on philosophical grounds, in favor of an untested, and mostly under-developed, alternative inferentialist semantic theory, which, at least in its present state, seems too crude to rigorously account even for fairly commonplace semantic phenomena. However, our discussion in §1 shows that this objection may be off target.

The objection conflates the levels of semantics and meta-semantics. The proponent of IRS can in principle be impressed with the advances of referential semantic theories and indeed endorse them, while maintaining that meta-semantic questions – MD and UND – call, in whole or in part, for an inferentialist treatment. In §2, we called inferentialists of this type orthodox. The inferentialist at the level of meta-semantics then faces the further question as to how the referentialist's semantic values are to interpreted: Are they to be taken at face value metaphysically or not? That is, in our terminology, is the inferentialist a *genuine* (orthodox) inferentialist or a *deflationary* one? As for MD, several authors (see, e.g., Peacocke, 1992; Hodes, 2004; MacFarlane, 2005) have maintained that an inferentialist

account of MD as applied to logical expressions is compatible with truth-conditional semantics for them – indeed, in suitable presentations of classical logic, inference rules *fix* the semantic interpretation of the logical vocabulary (e.g., Smiley, 1996; Rumfitt, 2000). The same is true for other expressions, such as deontic modals (see, e.g., Wedgwood, 2007; Chrisman, 2010) and indeed perhaps for the whole of language. Among these authors, some might advocate genuine variants (e.g., Wedgwood, 2007) others deflationary variants of deflationism (e.g., Brandom, 1994; 2000).

Williamson, however, remains skeptical. He writes:

> Since inferential relations do not fix truth and reference, meaning has not been adequately tied to the language-independent world. (Williamson, 2010, p. 23)

But Williamson simply *presupposes* the falsity of meta-semantic interpretations of MD. Pending an argument for this presupposition, it seems fair to conclude that the general worry against inferentialism has little to go on.

Let us turn, then, to the second general line of criticism. In a number of publications, Williamson has launched a sustained attack afflicting, according to him, various "programs which go under titles such as 'conceptual role semantics', 'inferentialism' and 'use theories of meaning'" (Williamson, 2006, p. 6, n. 5). Once again it is important to situate Williamson's objections within the map of inferentialist positions laid out in §2. The target of Williamson's criticism is the inferentialist's claim to be able to account for UND; it in no way concerns MD. What is more, Williamson's criticisms presuppose that understanding a sentence, for the inferentialist, is constituted by a grasp of certain epistemically analytic inferential relations. For instance, as we have seen, it may be a necessary condition in order to count as understanding 'and' that one appropriately recognizes the correctness of the inferential relation from $\ulcorner A$ and $B \urcorner$ to A and to B. The assumption, here, is that there must be some salient subset of the inferential relations an expression enters into that is constitutive of understanding it. Hence, on Williamson's picture, the inferentialist is committed to a version of epistemic analyticity in the sense of Boghossian (1996; 2003b). It is a consequence of this that, necessarily, someone who understands a given sentence, appropriately accepts certain inferences from and to that sentence. Williamson's objection to inferentialist theories of understanding do not, therefore, tell against theories of this type that do not rely on epistemic analyticity in this way. It is at least worth noting that not all inferentialist accounts of understanding are committed to epistemic analyticity in this way; our discussion of Brandom's holism in §3 is a case in point.

What, then, does Williamson's objection amount to? For simplicity, we summarize the objection as it is directed against any form of analyticity-based account of understanding; inferentialism amounts to a special case. His take-home message is this: understanding is an elastic notion – it cannot be adequately captured by any account that ties understanding to necessary conditions on the assent/dissent patterns of individual speakers or their acceptance of certain inferential relations. Consider the sentence

(`Vixen`) Every vixen is a female fox.

According to UND, necessarily, whoever understands `Vixen` assents to it. It is constitutive of one's understanding of '`Vixen`' that under suitable conditions (e.g., provided one

understands the remaining words occurring in the sentence), one assents to `Vixen`. Call a theory of understanding of this general form a *criterial* theory of understanding. Williamson provides a recipe for generating counter-examples to any such theory.

Suppose e is an expression, m its meaning. According to a criterial account of understanding, there must be a sentence or pattern of inference $C(e)$ such that a speaker's assent to it or appropriate recognition of it is necessary for her to count as understanding e. Williamson claims to have a general recipe for cooking up counter-examples to any such $C(e)$, for any e. Namely, we can imagine an expert on m who, on (possibly erroneous) theoretical grounds, rejects $C(e)$. By the criterialist's standards our expert does not understand e. But surely, by any ordinary standards, she *does* understand e – she is an expert, after all.

In order to get a better feel for the foregoing schematic objection, let us consider one of its instances. Suppose, as seems *prima facie* reasonable from the inferentialist's point of view, that an appropriate appreciation of *modus ponens* (MP) is constitutive of understanding 'if.' As for any case, Williamson believes a counter-example can be concocted also for this case. That is, a counter-example in which an expert whose semantic competence cannot reasonably be questioned rejects the criterial pattern of inference or sentence. Indeed, no concocting is even needed in this case, Williamson thinks: in Vann McGee we already have a ready-made, real-life example of someone who is undeniably an expert about conditionals but nevertheless denies the validity of MP. His denial is founded on a number well-known putative counter-examples involving nested conditionals (see, e.g., McGee, 1985, p. 462).

For present purposes, it does not matter if McGee is right – we may even suppose that his example is fallacious, and that 'if' in English satisfies MP after all.[28] What matters is that, his erroneous views about MP notwithstanding, McGee surely understands 'if.' All the same, according to UND it seems we must say that he does not. So UND must be false, or so Williamson argues.

Williamson constructs structurally similar cases for the material conditional, for 'for all' and for 'and' (Williamson, 2003; 2007; 2011; 2012). His conclusion is that agreement in understanding doesn't require perfect agreement in use. All that is needed for a speaker to understand an expression of e is that she fully participate in the social practice of using e within her linguistic community. He writes:

> Each individual uses words as words of a public language; their meanings are constitutively determined not individually but socially, through the spectrum of linguistic activity across the community as a whole. The social determination of meaning requires nothing like an exact match in use between different individuals; it requires only enough connection in use between them to form a social practice. Full participation in that practice constitutes full understanding. (Williamson, 2007, p. 91)

Williamson's argument, then, is premised on what is sometimes referred to as *social externalism* (Burge, 1979; 1986): linguistic understanding is always understanding of a public language whose meanings are typically non-individualistically individuated. On the assumption that the community-wide use of 'if' largely conforms to MP, it is this assumption that allows Williamson to maintain that McGee's understanding of 'if' is unaffected by his rejection of certain instances of the rule. How can the inferentialist respond?

We begin with two observations about the *scope* of the argument. To begin, the objection does not target IRS directly. Rather, it aims at undermining a certain account of

understanding that can be – and typically is – associated with it. What is more, the argument does not apply to holistic versions of IRS, such as Brandom's. According to Brandom's inferentialism, one's understanding of an expression *e* is constituted by one's grasp of the entire network of inferential connections in which *e* participates. But, since such a totality typically slightly varies from speaker to speaker, it may be argued that it is to be *expected*, on Brandom's view, that different speakers understand *e* equally well, and yet associate it with different inferential roles. That being said, however, semantic holism is a highly controversial view (Dummett, 1991; Pagin, 1997). Let us therefore consider possible lines of response in defense of non-holistic versions of UND.

Boghossian (2012a) has recently suggested the following inferentialist response. In his view, McGee understands 'if,' but he understands it differently. His deviant use simply shows that he attaches a different meaning to 'if.' As he puts it:

> All that the inferential role theorist is committed to saying is that, if [Vann] succeeds in altering his behavior with ['if'] and flouts a meaning-constituting rule […], then he necessarily means something different by 'if' […] It is better to call this "meaning change" rather than incompetence. (Boghossian, 2012a, p. 232)

As it stands, however, Boghossian's response is problematic. It is premised on an idiolectic version of MD (see 2012a, p. 233, fn. 2) which is flatly inconsistent with the social externalism assumed by Williamson's criticism. It may be that such idiolectic conceptions are ultimately more congenial to the inferentialist. Yet, it is an interesting question whether the inferentialist can respond to Williamson on his own terms. While we lack the space to develop a full response, we canvass the general shape such a response might take.

We begin with the observation that Boghossian's meaning-change approach is especially compelling in cases that are very similar to the ones Williamson considers. Consider, for instance, the case of a full-blooded intuitionist and of a classical logician – call them Michael and Tim, respectively – and let us assume that they are both highly proficient logicians. Arguably, Michael and Tim have a different understanding of 'not' (DeVidi and Solomon 2001; Dummett 2007; 2009).[29] It might be argued, then, that it would be problematic to insist, faced with such a difference, that intuitionist and classical logicians have the same *understanding* of 'not.' After all, intuitionists consciously use 'not' differently. For instance, unlike classical logicians, they refuse to assent to certain instances of the Law of Excluded Middle. It would not do justice to their semantic beliefs to insist that, in spite of their avowed intention to use 'not' according to its intuitionist meaning, 'not' means classical negation in their mouth. Intuitionists *reject* the classical meaning of 'not,' and *there is* a coherent, if arguably ultimately untenable, intuitionist meaning of 'not.' The same, it might be contended, applies to the McGee case.

But wouldn't this be, once again, inconsistent with Williamson's social externalism? Here it helps to observe that both Michael and Tim are, we are assuming, experts who have very close to full understanding. What makes them experts is that they are able to make authoritative pronouncements – what Burge (1986) calls 'normative characterizations' – regarding the criteria for correctly applying expressions related appropriately to their area of competence. They are in the business of *investigating and explicitly articulating the rules we ought to follow*. They are, in Kaplan's distinction, language 'creators' as opposed to mere 'consumers' (Kaplan, 1989a, p. 602). And since they can help shape a linguistic community's linguistic standards,

in view of their divergences in the use of 'not,' they may be plausibly viewed as belonging to different linguistic communities. It will not do to insist, as Williamson does, that small differences in use don't make a difference in understanding. In the present context, they do. It is only natural that the pronouncements of two experts be carefully examined. Even a small difference in the rules governing the use of an expression *e* is likely to imply a difference in that expression's content, on the natural assumption (shared by all parties) that such a content validates the rules for *e*'s correct use.

So much, then, for experts. What should the inferentialist say about Joe Shmoe's small but systematic deviance with respect to a given expression *e*? Wouldn't it be patronizing to claim that, because he doesn't conform to the community-wide use of *e*, he doesn't understand *e* (Williamson, 2003; 2007)? Here one should consider two cases. If Joe *defers to experts*, that is, if, for instance, he is disposed to be corrected by them, then it is open to the inferentialist – in keeping with Williamson's social externalism – to maintain that it is the experts' dispositions that determine *e*'s meaning, and that are constitutive of Joe's understanding of *e*. If, on the other hand, Joe does not defer to experts, and stubbornly insists in using *e* in a deviant way, then it would seem appropriate to say that Joe indeed has an idiosyncratic understanding of *e*: he understands *e* differently.

The inferentialist, then, might respond to Williamson's challenge by insisting that his examples all involve a discrepancy in use among *experts*. Williamson's key argumentative move, the inferentialist might diagnose, is to treat experts as *non-experts*, and to discard, *for this reason*, their idiosyncratic use of certain expressions as irrelevant for those expressions' meaning and understanding. Yet, by the social externalist's own lights, experts – language creators – who use a given expression in different ways, should be credited with *different understandings* of that expression.

If the foregoing considerations are along the right track, it would seem that Williamson's central objection to IRS misses its target after all. To be sure, much more remains to be said. For instance, the inferentialist who pursued a response to Williamson along these lines would have to explain how, for instance, rival logicians might be able to disagree about the validity of what would intuitively seem to be the same logical law. A disagreement about the Law of Excluded Middle, on the foregoing view, would no longer be about whether $A \vee \neg A$ is valid or not, but, say, about whether 'not' and 'or' in English should be interpreted classically or intuitionistically. The question whether this is a plausible feature of the view, or an unpalatable consequence, is, as far as we can see, still very much open.

Notes

1 Many thanks to Alex Miller for very helpful comments on a first draft of this paper. Over the years, we have greatly benefited from conversations we had with a number of friends and colleagues on inferentialism and related topics. We'd like to thank in particular Corine Besson, Paul Boghossian, Cesare Cozzo, Salvatore Florio, Dominic Gregory, Bob Hale, Ole T. Hjortland, Luca Incurvati, Rosanna Keefe, Hannes Leitgeb, Dag Prawitz, Stephen Read, Ian Rumfitt, Gil Sagi, Neil Tennant, Tim Williamson, and Crispin Wright.

2 Stalnaker (1997) uses the labels 'descriptive semantics' and 'foundational semantics' to make the same distinction. We follow the terminology in Kaplan (1989a), which by now is fairly well established. Brandom's labels 'formal semantics' and 'philosophical semantics' seem to designate the same distinction. See Brandom (1994, p. 143) and Weiss and Wanderer (2010, p. 342) for formulations of Brandom's distinction.

3 The level of thought is treated analogously:

(CD) *Content determination*. The contents of concepts are determined by their role in inference.

(GRA) *Grasping*. To grasp a concept is to know its role in inference.

4 See, e.g., Brandom's deflationary notions of reference and truth in his (1994), as well as the discussion between Dummett and Brandom (Weiss and Wanderer, 2010, chs 13 and 29).

5 That said, certain local forms of IRS do proclaim a (local) atomism. Neil Tennant (1987; 1997), for instance, argues for a form of atomistic IRS about logical concepts. On such a view, the meaning of any logical operator is independent of that of the other logical operators (indeed independent of any other expressions). Analogously, understanding a logical operator does not presuppose antecedent understanding of any other logical operators.

6 But see our discussion of Brandom's holism in § 3 below.

7 Block (1986), Loar (1981), Harman (1999), and Peacocke (1992) belong to this camp.

8 Proponents of normative versions of IRS include Boghossian (2003a), Brandom (1994), Dummett (1991), Wedgwood (2007), Whiting (2009).

9 See, e.g., Popper (1947, p. 220), Kneale (1956, pp. 254–255), and Dummett (1991, p. 247).

10 Dummett correctly observes that while

> it may […] be that the [representational] meanings of the logical constants [i.e., the truth-functions they denote] are *determined* by the logical laws that govern their use in deductive arguments […] this cannot be assumed – it needs to be *shown*. (Dummett, 1991, p. 205)

Carnap (1943) first showed that, in standard natural deduction systems, the rules for $\vee$, $\neg$, and $\rightarrow$ fail to determine their standard truth-conditions. Thus, it would seem, standard natural deduction systems are not hospitable at least to certain interpretations of MD. However, while this is sometimes thought to be a problem (see, e.g., Raatikainen, 2008), it need not be. For one thing, on certain assumptions, the rules still determine the standard *intuitionistic* meanings of $\vee$, $\neg$, and $\rightarrow$ (Garson, 2001). For another, in slightly less standard, though arguably equally 'natural,' natural deduction systems, the classical I- and E-rules for $\vee$, $\neg$, and $\rightarrow$ *do* determine their truth-conditions (see, e.g., Smiley, 1996; Rumfitt, 2000).

11 For instance, a connective $\odot$ satisfying the standard I-rules for $\wedge$ but only one of its E-rules would be intrinsically harmonious, and yet intuitively disharmonious: its E-rule would not allow us to infer from $A \odot B$ all that was required to introduce $A \odot B$ in the first place.

12 Right-to-left: suppose B follows from A. Since A also follows from $CG[A]$, B itself follows from $CG[A]$. Left-to-right: suppose B follows from $CG[A]$. Now assume A. By the Fundamental Assumption, $CG[A]$ itself follows. Hence, on our assumption that B follows from $CG[A]$, we may conclude B, as required.

13 Whether harmony is also a *sufficient* condition for logicality is a more delicate question. See Read (2000).

14 See, also e.g., Hacking (1979, pp. 296) and Dummett (1991, pp. 217–218), and the discussion in Steinberger (2011a).

15 Alternatively, harmonious axiomatizations of classical logic can be given once *multiple conclusions* are allowed (Read, 2000; Cook, 2005), either in a natural deduction or in a sequent-calculus setting. Sequent calculi axiomatizations of intuitionistic and classical logic are exactly alike, except that classical sequent calculi allow for sequents with multiple premises *and* multiple conclusions. In turn, such sequents can be plausibly interpreted as saying that one may not assert all the antecedents and deny all the succedents, where, again, assertion and denial are both primitive speech acts (Restall, 2005). For a technical introduction to multiple-conclusion logics, see Shoesmith and Smiley (1978). For a recent criticism, see Steinberger (2011b).

16 We should stress that Brandom is but one proponent of a global form of inferentialism. For some representative publications on various global brands of IRS, see, e.g., Sellars (1956), Harman (1987), Field (1977; 1994; 2001), Block (1987; 1998), Cozzo (1994), Horwich (1998), Peacocke (1992), Boghossian (2003a; 2003b; 2012a; 2012b), Whiting (2006; 2008; 2009), Chalmers (2014). For reasons of space, we only focus on Brandom's work, on account of its influence, and of the detail in which the views have been worked out over the years.

17 For a criticism of Brandom's deflationary approach see Shapiro (2004). See MacFarlane (2010) for an interesting objection to Brandom's assumption that a use-theoretic approach is incompatible with a truth-conditional approach.

18 As we noted in section §2, there will be extremal cases – language-entry rules – in which the grounds for inferring a sentence may be perceptual; and in which the consequences of asserting a sentences will be non-linguistic.

19 Brandom explicitly follows Sellars on this point Brandom (1994, p. 97).

20 See in particular Brandom's discussion of Dummett's inferentialist treatment of pejoratives (Brandom, 2000, pp. 69ff.). See Williamson (2009) for a stern referentialist reprisal.

21 The exchange between Hale and Wright and Brandom is illuminating in this regard. See Weiss and Wanderer (2010, chs 17 and 33).

22 For an impression of the breadth of criticisms inspired by his (Brandom, 1994) alone, readers may consult, e.g., the book symposium in a special issue of *Philosophy and Phenomenological Research* (vol. 57, no. 1, 1997), and Weiss and Wanderer (2010).

23 For more discussion on semantic holism, see, e.g., Dummett (1991), Harman (1993), Pagin (1997; 2009), Cozzo (2002).

24 See Brandom (1994, p. 666) for more detailed discussion.

25 For an incomplete sample, see, e.g., Fodor and Lepore (1991), Casalegno (2004), Williamson (2003; 2007; 2009; 2012), Horwich (2005), Besson (2012), Schechter and Enoch (2006), Dogramaci (2012). For some inferentialist responses beyond the ones cited below, see, e.g., Eklund (2007), Balcerak Jackson (2009), and Murzi and Steinberger (2013).

26 The same point is also forcefully made, *inter alia*, in Jason Stanley's introduction to Stanley (2007).

27 See also Williamson (2007, p. 282).

28 If McGee is right, and 'if' actually does not satisfy MP, contrary to what we're assuming, then expert speakers who infer according to the unrestricted rule of MP would serve as purported counter-examples to UND.

29 Intuitionist and classical negation (respectively, $\sim$ and $\neg$) have different meanings on most intuitionist semantics. For instance, on the standard BHK (Brouwer–Heyting–Kolmogorov) semantics $\sim A$ means that there is no proof of A, i.e., that it is *impossible* to prove A. Given the intuitionist's equation of truth with the existence of a proof, the impossibility of proving A in turn entails that A *can't* be true. Similarly, on the standard Kripke semantics for intuitionist logic, $\sim A$ is forced by a state of information w if and only if *no possible development* of w forces A. This modal component is arguably absent in classical negation (DeVidi and Solomon, 2001; Dummett 2007; 2009).

References

Balcerak Jackson, B. 2009. "Understanding and semantic structure: Reply to Timothy Williamson." *Proceedings of the Aristotelian Society*, 109(1): 337–343. DOI:10.1111/j.1467-9264.2009.00272.x.

Belnap, N. 1962. "Tonk, plonk and plink." *Analysis*, 22(6): 130–134.

Besson, C. 2012. "Logical knowledge and ordinary reasoning." *Philosophical Studies*, 158(1): 59–82. DOI:10.1007/s11098-010-9672-3.

Block, N. 1986. "Advertisement for a semantics for psychology." *Midwest Studies in Philosophy*, 10(1): 615–678. DOI:10.1111/j.1475-4975.1987.tb00558.x.

Block, N. 1987. "Functional role and truth conditions." *Proceedings of the Aristotelian Society*, 61: 157–181.

Block, N. 1998. "Conceptual role semantics." In *Routledge Encyclopedia of Philosophy*, edited by E. Craig, pp. 242–256. London: Routledge.

Boghossian, P. 1996. "Analyticity reconsidered." *Noûs*, 30(3): 360–391. DOI:10.2307/2216275.

Boghossian, P. 2003a. "Blind reasoning." *Proceedings of the Aristotelian Society*, 77(1): 225–248. DOI:10.1111/1467-8349.00110.

Boghossian, P. 2003b. "Epistemic analyticity: a defense." *Grazer Philosophische Studien*, 66: 15–35.

Boghossian, P. 2011. "Williamson on the *a priori* and the analytic." *Philosophy and Phenomenological Research*, 82(2): 488–497. DOI:10.1111/j.1933-1592.2010.00395.x.

Boghossian, P. 2012a. "Inferentialism and the epistemology of logic: reflections on Casalegno and Williamson." *Dialectica*, 66(2): 221–236. DOI:10.1111/j.1746-8361.2012.01303.x.

Boghossian, P. 2012b. "What is inference?" *Philosophical Studies*, 169(1): 1–18. DOI:10.1007/s11098-012-9903-x.

Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Brandom, R. 2007. "Inferentialism and some of its challenges." *Philosophy and Phenomenological Research*, 74(3): 651–676. DOI:10.1111/j.1933-1592.2007.00044.x.

Brandom, R. 2008. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: Oxford University Press.

Burge, T. 1979. "Individualism and the mental." *Midwest Studies in Philosophy*, 4(1): 73–121. DOI:10.1111/j.1475-4975.1979.tb00374.x.

Burge, T. 1986. "Intellectual norms and foundations of mind." *The Journal of Philosophy*, 85: 649–663. DOI:jphil198683121.

Carnap, R. 1943. *Formalization of Logic*. Cambridge, MA: Harvard University Press.

Casalegno, P. 2004. "Logical concepts and logical inferences." *Dialectica*, 58(3): 395–411.

Chalmers, D. 2014. *Constructing the World*. Oxford: Oxford University Press.

Chrisman, M. 2010. "From epistemic expressivism to epistemic inferentialism." In *Social Epistemology*, edited by A. Haddock, A. Millar, and D. Pritchard, pp. 112–128. Oxford: Oxford University Press.

Chrisman, M. 2015. "Metanormative theory and the meaning of deontic modals." Unpublished manuscript.

Cook, R. 2005. "Intuitionism reconsidered." In *The Oxford Handbook for Philosophy of Mathematics and Logic*, edited by S. Shapiro, pp. 387–411. Oxford: Oxford University Press.

Cozzo, C. 1994. "What can we learn from the paradox of knowability?" *Topoi*, 13(2): 71–78. DOI:10.1007/BF00763505.

Cozzo, C. 2002. "Does epistemological-holism lead to meaning-holism?" *Topoi*, 21(1–2): 25–45. DOI:10.1023/A:1014876214057.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

De Vidi, D., and G. Solomon. 2001. "Knowability and intuitionistic logic." *Philosophia*, 28(1): 319–334. DOI:10.1007/BF02379783.

Dogramaci, S. 2012. "Apriority." In *The Routledge Companion to Philosophy of Language*, edited by D. Graff-Fara and G. Russell, pp. 768–781. London: Routledge.

Dretske, F. 2000. *Perception, Knowledge and Belief*. Cambridge: Cambridge University Press.

Dummett, M. 1973. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Dummett, M. 1993. "Wittgenstein on necessity: some reflections." In *The Seas of Language*. Cambridge, MA: Harvard University Press.

Dummett, M. 2007. "Reply to Wolfgang Künne." In *The Philosophy of Michael Dummett*, edited by R. Auxier and L. Hahn, pp. 345–50. Chicago: Open Court.

Dummett, M. 2009. "Fitch's paradox of knowability." In *New Essays on the Knowability Paradox*, edited by J. Salerno, pp. 51–52. Oxford: Oxford University Press.

Eklund, M. 2007. "Meaning-constitutivity." *Inquiry*, 50(6): 559–574. DOI:10.1080/00201740701698506.

Field, H. 1977. "Logic, meaning, and conceptual role." *Journal of Philosophy*, 74(7): 379–409. DOI:10.2307/2025580.

Field, H. 1994. "Are our logical and mathematical concepts highly indeterminate?" *Midwest Studies in Philosophy*, 19(1): 391–429. DOI:10.1111/j.1475-4975.1994.tb00296.x.

Field, H. 2001. "Attributions of meaning and content." In *Truth and the Absence of Fact*, pp. 157–174. Oxford: Clarendon Press.

Fodor, J. 1990. *A Theory of Content*. Cambridge, MA: MIT Press.

Fodor, J., and E. Lepore. 1991. "Why meaning (probably) isn't conceptual role." *Mind & Language*, 6(4): 328–343. DOI:10.1111/j.1468-0017.1991.tb00260.x.

Fodor, J., and E. Lepore. 2001. "Brandom's burdens: compositionality and inferentialism." *Philosophy and Phenomenological Research*, 63(2): 465–481. DOI:10.2307/3071079.

Fodor, J., and E. Lepore. 2010. "Brandom beleaguered". In *Reading Brandom*, edited by B. Weiss and J. Wanderer, pp. 181–194. Abingdon: Routledge.

Garson, J. 2001. "Natural semantics: why natural deduction is intuitionistic." *Theoria*, 67(2): 114–137. DOI:10.1111/j.1755-2567.2001.tb00200.x.

Gentzen, G. 1934. "Untersuchungen über das logischen Schließen." *Mathematische Zeitschrift*, 39: 405–431.

Greenberg, M., and G. Harman. 2006. "Conceptual role semantics." In *The Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. Smith, pp. 295–322. Oxford: Oxford University Press.

Grice, P. 1957. "Meaning." *Philosophical Review*, 66(3): 377–388.

Hacking, I. 1979. "What is logic?" *Journal of Philosophy*, 76(6): 285–319. DOI:10.2307/2025471.

Harman, G. 1987. "(Nonsolipsistic) conceptual role semantics." In *New Directions in Semantics*, edited by Ernest Lepore, pp. 242–256. Waltham: Academic Press.

Harman, G. 1993. "Meaning holism defended." *Grazer Philosophische Studien*, 46: 163–171. DOI:10.5840/gps1993467.

Harman, G. 1999. *Reasoning, Meaning, and Mind*. Oxford: Oxford University Press.

Hodes, H. 2004. "On the sense and reference of a logical constant." *The Philosophical Quarterly*, 54(214): 134–165. DOI:10.1111/j.0031-8094.2004.00345.x.

Horwich, P. 1998. *Meaning*. Oxford: Oxford University Press.

Horwich, P. 2005. *Reflections on Meaning*. Oxford: Oxford University Press.

Incurvati, L., and P. Smith. 2010. "Rejection and valuations." *Analysis*, 70(1): 3–10. DOI:10.1093/analys/anp134.

Kaplan, D. 1989a. "Afterthoughts." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 565–614. Oxford: Oxford University Press.

Kaplan, D. 1989b. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Oxford University Press.

Kneale, W. 1956. "The province of logic." In *Contemporary British Philosophy*, edited by H. D. Lewis, pp. 237–261. London: George Allen & Unwin.

Loar, B. 1981. *Mind & Meaning*. Cambridge: Cambridge University Press.

MacFarlane, J. 2005. "Logical constants." In *Stanford Encyclopedia of Philosophy* (Fall 2015 edn), edited by E. Zalta. http://plato.stanford.edu/archives/fall2015/entries/logical-constants (accessed August 18, 2016).

MacFarlane, J. 2010. "Pragmatism and inferentialism." In *Reading Brandom*, edited by B. Weiss and J. Wanderer, pp. 81–96. Abingdon: Routledge.

McGee, V. 1985. "A counterexample to modus ponens." *The Journal of Philosophy*. 82(9): 462–471. DOI:jphil198582937.

Murzi, J., and F. Steinberger. 2013. "Is knowledge of logic dispositional?" *Philosophical Studies*, 166(1): 165–183. DOI:10.1007/s11098-012-0063-9.

Pagin, P. 1997. "Is compositionality compatible with holism?" *Mind & Language*, 12(1): 11–33. DOI:10.1111/1468-0017.00034.

Pagin, P. 2009. "Intuitionism and the anti-justification of bivalence." In *Logicism, Intuitionism, and Formalism: What has Become of Them?* edited by S. Lindström *et al*. Dordrecht, Netherlands: Synthese Library, Springer.

Parsons, T. 1984. "Assertion, denial and the liar paradox." *Journal of Philosophical Logic*, 13(2): 136–152. DOI:10.1007/BF00453019.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Popper, K. 1947. "New foundations for logic." *Mind*, 56(223): 193–235. DOI:10.1093/mind/LVI.223.193.

Prawitz, D. 1965. *Natural Deduction*. Stockholm: Almqvist and Wiksell.

Prawitz, D. 1974. "On the idea of a general proof theory." *Synthese*, 27(1): 63–77. DOI:10.1007/BF00660889.

Prawitz, D. 1977. "Meaning and proofs: on the conflict between classical and intuitionistic logic." *Theoria*, 43(1): 2–40. DOI:10.1111/j.1755-2567.1977.tb00776.x.

Prawitz, D. 1985. "Remarks on some approaches to the concept of logical consequence." *Synthese*, 62(2): 153–171. DOI:10.1007/BF00486044.

Prior, A. 1960. "The runabout inference-ticket." *Analysis*, 21(2): 38–39. DOI:10.1093/analys/21.2.38.

Putnam, H. 1975. "The meaning of meaning." In *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, H. 1981. *Realism, Truth and History*. Cambridge: Cambridge University Press.

Raatikainen, P. 2008. "On rules of inference and the meanings of logical constants." *Analysis*, 68(4): 282–287. DOI:10.1111/j.1467-8284.2008.00754.x.

Read, S. 2000. "Harmony and autonomy in classical logic." *Journal of Philosophical Logic*, 29(2): 123–154. DOI:10.1023/A:1004787622057.

Reiss, J. 2012. "Causation in the sciences: an inferentialist account." *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 43(4): 769–777. DOI:10.1016/j.shpsc.2012.05.005.

Restall, G. 2005. "Multiple conclusions." In *Logic, Methodology and the Philosophy of Science: Proceedings of the Twelfth International Congress*, edited by P. Hájek, L. Valdés-Villanueva, and D. Westertsthål, pp. 189–205. London: King's College Publications.

Rosen, G. 1997. "Who makes the rules around here?" *Philosophy and Phenomenological Research*, 57(1): 163–171. DOI:10.2307/2953786.

Rumfitt, I. 2000. "'Yes' and 'No.'" *Mind*, 109(436): 781–824. DOI:10.1093/mind/109.436.781.

Russell, G. 2014. "Metaphysical analyticity and the epistemology of logic." *Philosophical Studies*, 171(1): 161–175. DOI:10.1007/s11098-013-0255-y.

Salerno, J., ed. 2009. *New Essays on the Knowability Paradox*. Oxford: Oxford University Press.

Schechter, J., and D. Enoch. 2006. "Meaning and justification: the case of modus ponens." *Noûs*, 40(4): 687–715. DOI:10.1111/j.1468-0068.2006.00629.x.

Sellars, W. 1953. "Inference and meaning." *Mind*, 62(247): 313–338. DOI:10.1093/mind/LXII.247.313.

Sellars, W. 1956. "Empiricism and the philosophy of mind." *Minnesota Studies in The Philosophy of Science: The Foundations of Science and the Concepts of Psychology and Psychoanalysis I*, edited by H. Feigl and M. Scriven, pp. 253–329. Minneapolis: University of Minnesota Press.

Shapiro, L. 2004. "Brandom on the normativity of meaning." *Philosophy and Phenomenological Research*, 68(1): 141–160. DOI:10.1111/j.1933-1592.2004.tb00330.x.

Shoesmith, D. J., and T. Smiley. 1978. *Multiple-Conclusion Logic*. Cambridge: Cambridge University Press.

Smiley, T. 1996. "Rejection." *Analysis*, 56(1): 1–9. DOI:10.1111/j.0003-2638.1996.00001.x.

Stalnaker, R. 1997. "Reference and necessity." In *A Companion to the Philosophy of Language*, edited by B. Hale and C. Wright, pp. 534–554. Oxford: Blackwell.

Stanley, J. 2007. *Language in Context*. Oxford: Oxford University Press.

Steinberger, F. 2009. "Harmony and Logical Inferentialism." PhD diss., University of Cambridge.

Steinberger, F. 2011a. "What harmony could and could not be." *Australasian Journal of Philosophy*, 89(4): 617–639. DOI:10.1080/00048402.2010.528781.

Steinberger, F. 2011b. "Why conclusions should remain single." *Journal of Philosophical Logic*, 40(3): 333–355. DOI:10.2307/41487518.

Suarez, M. 2004. "An inferential conception of scientific representation." *Philosophy of Science*, 71(5): 767–779. DOI:10.1086/421415.

Tennant, N. 1987. *Anti-realism and Logic*. Oxford: Clarendon Press.
Tennant, N. 1997. *The Taming of the True*. Oxford: Oxford University Press.
Tennant, N. 2008. "Inferentialism, logicism, harmony, and a counterpoint." In *Essays for Crispin Wright: Logic, Language and Mathematics*, edited by A. Coliva. Oxford: Oxford University Press.
Warren, J. 2015. "The possibility of truth by convention." *The Philosophical Quarterly*, 65(258): 84–93. DOI:10.1093/pq/pqu051.
Wedgwood, R. 2007. *The Nature of Normativity*. Oxford: Oxford University Press.
Weiss, B., and J. Wanderer, eds. 2010. *Reading Brandom*. Abingdon: Routledge.
Whiting, D. 2006. "Conceptual role semantics." *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/conc-rol/ (accessed August 18, 2016).
Whiting, D. 2008. "Meaning holism and *de re* ascription." *Canadian Journal of Philosophy*, 38(4): 575–599. DOI:10.1353/cjp.0.0033.
Whiting, D. 2009. "On epistemic conceptions of meaning: use, meaning and normativity." *European Journal of Philosophy*, 17(3): 416–434. DOI:10.1111/j.1468-0378.2008.00320.x.
Williamson, T. 2003. "Blind reasoning: understanding and inference." *Proceedings of the Aristotelian Society*, 77(1): 249–293. DOI:10.1111/1467-8349.00111.
Williamson, T. 2006. "Conceptual truth." *Proceedings of the Aristotelian Society*, 80(1): 1–41. DOI:10.1111/j.1467-8349.2006.00136.x.
Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Oxford University Press.
Williamson, T. 2009. "Reference, inference and the semantics of pejoratives." In *The Philosophy of David Kaplan*, edited by J. Almog and P. Leonardi, pp. 137–158. Oxford: Oxford University Press.
Williamson, T. 2010. "Review of Robert Brandom, *Reason in Philosophy: Animating Ideas*." *Times Literary Supplement*, 5579: 22–23.
Williamson, T. 2011. "Reply to Boghossian." *Philosophy and Phenomenological Research*, 82(2): 498–506. DOI:10.1111/j.1933-1592.2010.00400.x.
Williamson, T. 2012. "Boghossian and Casalegno on understanding and inference." *Dialectica*, 66(2): 237–247. DOI:10.1111/j.1746-8361.2012.01295.x.

Further Reading

Brandom, R. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press. An accessible introduction to the most worked out global inferentialist position.
Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press. A difficult but rewarding exploration of semantic antirealism. Chapter 11, in particular, has proved influential for the program of logical inferentialism.
Tennant, N. 1997. *The Taming of the True*. Oxford: Oxford University Press. A carefully argued book-length treatment of the motivations for and the ramifications of (logical) inferentialism.
Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Oxford University Press. An important book on the methodology of philosophy. Chapter 4 contains challenging objections to inferentialism and use theories of meaning in general.

10

Against Harmony

IAN RUMFITT[1]

In both natural deduction and sequent formalizations of logical systems, each connective is associated with an introduction rule and an elimination rule. The introduction rule for the connective **C** is one which licenses the derivation *of* a formula dominated by **C**; the elimination rule is one which licenses the deduction of further conclusions *from* such a formula, often with other formulae as auxiliary premises. In the sense of the term with which I shall be concerned, *harmony* is a particular relationship between the introduction rule and the elimination rule for a given connective. Whether the rules of a logical system are harmonious is certainly of great interest to proof theorists, but I am concerned with a philosophical claim about the notion. The *Harmony Thesis*, as I shall call it, says that a connective is defective unless its associated introduction and elimination rules are in harmony. It also says that a connective is defective if the logical principles which regulate its use go beyond a pair of harmonious introduction and elimination rules. Most proponents of the Harmony Thesis have, indeed, a particular defect in mind. On their view, a connective will not possess a sense – it will not have a coherent meaning – unless its logical behavior is regulated only by a pair of harmonious introduction and elimination rules.

The Harmony Thesis is connected to wider claims in the philosophy of language and the philosophy of logic. According to *Inferential Role Semantics* (IRS), the meaning of a complete statement is determined by its role in inference, and the meanings of sub-sentential expressions are determined by their contribution to the inferential roles of complete statements in which they figure. As Julien Murzi and Florian Steinberger remark in their contribution to this volume (see Chapter 9, INFERENTIALISM), many adherents of IRS appeal to the Harmony Thesis in order to circumscribe the range of meaning-determining inferential roles. We have yet to see what harmony comes to, but it is also widely held that the classical introduction and elimination rules for negation violate the Thesis, so Harmony is often invoked in challenges to classical logic. Dummett's *The Logical Basis of Metaphysics* provides a famous example of this. He holds that a connective's introduction and elimination rules must be in harmony if the connective is to make sense. So he takes the perceived lack

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

of harmony between the classical rules for negation to be "a strong ground for suspicion that the supposed understanding [of classical negation] is spurious" (Dummett, 1991, p. 299).

In this chapter, I want to scrutinize the most influential arguments which have been put forward for the Harmony Thesis. I find these arguments wanting, so my conclusion will be that the Thesis is not well supported. Any rejection of it must be provisional: tomorrow, someone may come up with a brilliant new argument which will persuade us all that there is a requirement of harmony on any intelligible connective. But I doubt it. As the analysis will reveal, the most popular arguments for the Harmony Thesis are not near misses which might succeed with a bit of tweaking. Rather, they are superficially plausible arguments which turn out, on closer examination, to rely upon highly dubious premises from epistemology and the theory of meaning. My analysis will not challenge the claim that deductive systems exhibiting harmony have attractive proof-theoretical features; to the contrary, I regard that claim as obviously true. But a connective may possess a sense, and may in other ways be non-defective, without generating those nice features, so the proof-theoretic elegance of harmonious rules does not settle the philosophical questions I am addressing. I conclude by discussing briefly how the failure of the Harmony Thesis affects the prospects for IRS.

1 The Inversion Principle

I have been using the term 'harmony,' but what exactly does it mean? As we shall see, different parties mean rather different things by it. I start by expounding what I take to be the most prominent version of the Harmony Thesis.

Central to this version is the claim that a connective's introduction rule determines its sense or meaning. The claim goes right back to Gentzen, who wrote that "the introduction rules represent, so to say, the 'definitions' of the signs in question [i.e., the connectives and quantifiers], and the elimination rules are, in the final analysis, no more than consequences of these definitions … In eliminating a sign, we may use a formula which has that sign as its main connective only with the meaning afforded it by the sign's introduction rule" (Gentzen, 1935, p. 189/1969, p. 80). Prawitz (see especially 1974), Dummett (see 1991), and Negri and von Plato (2001) also accept this claim.

What, exactly, are introduction rules and elimination rules? As we shall see, some delicate issues surround this question, but an initial answer runs as follows. While they may be extended to natural languages, the notions were originally applied to formalized languages, so let us consider for simplicity a propositional language L with sentence letters P_1, P_2, …, and a collection of finitary connectives. A *sequent* is then defined to be an expression $X \Rightarrow A$, where A is a formula of L and X is a finite set (possibly empty) of such formulae.[2] Where $n \geq 0$, consider the $(n+1)$-tuple of sequents $<X_1 \Rightarrow A_1, \ldots, X_n \Rightarrow A_n, Y \Rightarrow B>$. This determines an $(n+1)$-ary *rule for sequents* (for short, a *rule*), comprising all substitution instances of the $(n+1)$-tuple. This rule is understood as licensing the passage from any substitution instance of the first n sequents (the *premise sequents*) to the corresponding instance of the last sequent (the *conclusion sequent*). We mark the division between the premise and conclusion sequents with the solidus, /. When $n = 0$, we have a *rule of inference*, such as $\wedge$-introduction:

$$/P_1, P_2 \Rightarrow P_1 \wedge P_2$$

(when displaying particular rules, I shall often omit the angled brackets around tuples). When $n \geq 1$, we have a *rule of proof*, such as $\rightarrow$-introduction:

$$X, P_1 \Rightarrow P_2 / X \Rightarrow P_1 \rightarrow P_2.$$

A rule is *elementary* if all its members are substitution instances of a tuple $X_1 \Rightarrow A_1$, …, $X_n \Rightarrow A_n / Y \Rightarrow B$ such that (a) all the formulae in all the sequents are sentence letters except for at most one, which is of the form $\mathbf{C}(P_1, \ldots, P_k)$ for some k-place connective **C**; and (b) this complex formula $\mathbf{C}(P_1, \ldots, P_k)$ (if it appears at all) occurs either on the left or the right of the conclusion sequent $Y \Rightarrow B$. If it occurs on the right of $\Rightarrow$, the elementary rule in question is an *introduction rule* for **C**. If it occurs on the left, it is an *elimination rule* for **C**.[3]

On this account, the rule of double negation elimination (*DNE*) – from $\neg\neg A$, infer A – does not qualify as an elimination rule. While this consequence may initially seem surprising, it is to be welcomed. Gentzen described *DNE* as an elimination rule: *DNE*, he wrote, "represents a new elimination of negation whose admissibility does not follow at all from our way of introducing the negation sign by the ¬-*I* rule" (Gentzen, 1935, p. 190/1969, p. 81). Dummett followed him in this. Indeed, Dummett's critique of classical logic in *The Logical Basis of Metaphysics* is really an extended elaboration of the sentence just quoted from Gentzen: the classical elimination rule for negation, *viz. DNE*, is not in harmony with its introduction rule. But *DNE* does not conform to Dummett's own account of what an elimination rule is. In applying *DNE* we pass, of course, to a conclusion that contains two fewer occurrences of '¬' than does the premise, but that is not enough for it to count as an elimination rule. "In the case of a logical constant, we may regard the introduction rules governing it as giving the conditions for the assertion of a statement of which it is the main operator, and the elimination rules as *giving the consequences of such a statement*" (Dummett, 1981, pp. 454–455; emphasis added). The *DNE* rule tells us nothing *in general* about the consequences of statements in the form $\ulcorner\neg A\urcorner$. It tells us something only about the very special case of statements in the form $\ulcorner\neg\neg A\urcorner$. It is to the good, then, that the proposed account does not classify *DNE* as an elimination rule.

In what sense might an introduction rule for **C** be thought to *define* **C**? Gentzen did not say but, according to Prawitz and Dummett, it does so by specifying the *direct grounds* (alias the 'canonical' grounds) for asserting a formula dominated by **C**. Suppose that G_1 is a direct ground for asserting the interpreted formula A and that G_2 is a direct ground for asserting the interpreted formula B. Then the rule of $\wedge$-introduction tells us that the combination of G_1 with G_2 constitutes a direct ground for asserting the conjunctive formula $\ulcorner A \wedge B\urcorner$. Indeed, if this rule is to *define* the sense of '$\wedge$,' it must be understood as telling us that the *only* direct grounds for asserting $\ulcorner A \wedge B\urcorner$ are those which combine a direct ground for A with a direct ground for B. The introduction rule for '$\vee$' is $<P_1 / P_1 \vee P_2> \cup <P_2 / P_1 \vee P_2>$. In a similar way, this rule is to be read as telling us that a direct ground for asserting a disjunctive formula $\ulcorner A \vee B\urcorner$ will be either a direct ground for A or a direct ground for B. As remarked, the rule of $\rightarrow$-introduction is a rule of proof, not a rule of inference, so here matters are less straightforward. But $\rightarrow$-introduction is understood as telling us that a direct ground for asserting $\ulcorner A \rightarrow B\urcorner$ is a method for transforming any ground for A into a ground for B. There are of course grounds for assertion – indeed, conclusive grounds for assertion – which are not direct. Thus we might assert $\ulcorner A \wedge B\urcorner$, not on the basis of the combination of G_1 with G_2, but on the strength of a deduction of $\ulcorner A \wedge B\urcorner$ from the premises C and $\ulcorner C \rightarrow A \wedge B\urcorner$.

Any development of this theory of direct or canonical grounds clearly faces problems. For one thing, the method that constitutes a direct ground for asserting $\ulcorner A \rightarrow B\urcorner$ needs to be one that transforms *any* ground for A into a ground for B, so the specification of direct grounds would appear not to be straightforwardly compositional.[4] It is, however, in the context of this conception of the meaning of the connectives that the present version of the Harmony Thesis belongs. For suppose that the meaning of a connective **C** is given by its introduction rule; then the elimination rule for **C** must be faithful to that meaning.

In Gentzen's words, we may use **C** only with the meaning that the introduction rule affords it. On this view, the requirement of harmony does no more than spell out what such fidelity consists in. Gentzen conveys the requirement he has in mind only by way of an example: "if we wished to use [the formula $\ulcorner A \rightarrow B \urcorner$] by eliminating the $\rightarrow$-symbol … we could do this precisely by inferring B directly, once A has been proved, for what $\ulcorner A \rightarrow B \urcorner$ attests is just the existence of a derivation of B from A" (Gentzen, 1935, p. 189/1969, pp. 80–81, with incidental changes in the symbolism). Negri and von Plato, though, spell out the general principle to which Gentzen implicitly appeals. To find the elimination rule which is faithful to a given introduction rule, they write, "we ask what the conditions are, in addition to assuming the major premiss derived, that are needed to satisfy the *Inversion Principle*:

> Whatever follows from the direct grounds for deriving a formula must follow from that formula. (Negri and von Plato, 2001, p. 6; I write 'formula' where they have 'proposition')

According to the present version of the Harmony Thesis, then, a non-defective logical connective must be regulated only by a pair of introduction and elimination rules which satisfy the Inversion Principle. This version of the Thesis is justified by the claim that the elimination rule for a connective must be faithful to the introduction rule that defines the connective's meaning. One finds similar, albeit less explicit, formulations of this version of the Thesis, and of the suggested justification for it, in Prawitz and Dummett.[5]

One merit which Negri and von Plato claim for their formulation is that it not only justifies a certain elimination rule, given an introduction rule, but "actually determines what the elimination rules corresponding to given introduction rules should be" (2001, p. xvi).[6] Take disjunction as an example. The direct grounds for $\ulcorner A \vee B \urcorner$, we saw, are either direct grounds for A or direct grounds for B. The Inversion Principle is understood to say that whatever follows from *any* of the direct grounds for asserting a formula must follow from that formula. So we reach the $\vee$-elimination rule in the form: given a derivation of C from the assumption A, and another derivation of it from the assumption B, we may derive C from the disjunction $\ulcorner A \vee B \urcorner$. This is, in fact, the form of $\vee$-elimination that Negri and von Plato take their Inversion Principle to yield (2001, p. 7) and they go on to show how to excise from a derivation any deductive steps in which an instance of $\vee$-introduction is immediately followed by an instance of $\vee$-elimination. Suppose, for example, that we apply $\vee$-introduction to derive $\ulcorner A \vee B \urcorner$ from A, and then immediately eliminate $\ulcorner A \vee B \urcorner$ to reach the conclusion C. The derivation will then have the form

$$\dfrac{\begin{array}{ccc}\Delta & A & B \\ A & \Delta' & \Delta'' \\ A \vee B & C & C\end{array}}{C}$$

and we may simplify it by cutting out the occurrence of $\ulcorner A \vee B \urcorner$ entirely, thereby reaching

$$\begin{array}{c}\Delta \\ A \\ \Delta' \\ C\end{array}$$

This is an example of what Prawitz labels a 'reduction step' and of what Dummett calls 'levelling a local peak.'

There is, however, a problem here. The form of ∨-elimination that Negri and von Plato's Inversion Principle yields is the restricted version of the rule found in quantum logic, in which the conclusion *C* must be derived from *A*, and from *B*, without the use of any side premises.[7] However, only a few pages later in their treatise on proof theory (2001, p. 15), they blithely reformulate the elimination rule in the form *with* side premises: it is this stronger form of the rule that is found in classical, intuitionistic, and indeed minimal logic. It is hard to see what justifies the switch: in identifying the elimination rule that matches a given introduction rule, Negri and von Plato tell us, we are to "ask what the conditions are … that are *needed* to satisfy the Inversion Principle" (2001, p. 6, with emphasis added): the 'needed' seems to imply that we are to select the *weakest* elimination rule which satisfies Inversion.[8] I shall return to this problem in due course.

2 An Argument for the Inversion Principle

First, though, we should consider the central question which this attempted justification of the Harmony Thesis raises: why should we accept the Inversion Principle?

At first blush, there seems to be a compelling argument for the Principle. As we have seen, it is to be read as saying: 'Whatever follows from *any* of the direct grounds for asserting a formula must follow from that formula.' Let *C* be a formula which follows from any of the direct grounds for asserting a formula *A*, and suppose that *A* is asserted. If *A* has been correctly asserted, one might think, then at least one of its direct grounds must obtain. *Ex hypothesi*, *C* follows from any such ground. So *C* must obtain if *A* has been correctly asserted. *C*, then, is a commitment of a correct assertion of *A*, and as such – one might think – *C* must follow from *A*. So far from being compelling, however, this argument faces two severe problems.

First, and most obviously, the argument implicitly rejects the view that consequence is a matter of the preservation (or necessary preservation) of *truth* in favor of a view whereby consequence is a matter of the preservation (or necessary preservation) of correct *assertibility*. For suppose one did think that consequence was a matter of the preservation of truth. In that case, the proposed justification of Harmony would scarcely get started. On this view, in order to argue that *C* follows from *A*, we would need to show that *C* is true given only the assumption that *A* is *true*. From the assumption that *A* is true, however, it does not follow that any direct ground for asserting *A* obtains. Indeed, it does not follow that *any* ground for asserting *A* obtains. For all that has been said, the formula *A* might be true but unassertible. Only if consequence is understood to consist in the preservation of correct assertibility, then, does the mooted argument so much as get going. On that alternative view of consequence, in seeking to show that *C* follows from *A*, we *shall* start by assuming that *A* is correctly asserted.

A second problem confronts the argument, though, even if we do take consequence to consist in the preservation of correct assertibility rather than in the preservation of truth. On this view, in trying to show that *C* follows from *A* we shall start by assuming that *A* is correctly asserted, from which it follows that some ground for asserting *A* obtains. What we are given about *C*, though, is that it follows from any *direct* ground for asserting *A*. So in order to conclude from our assumption that *C* obtains, we shall need a premise to the effect that whenever a ground for asserting a formula obtains, some *direct* ground for asserting it obtains. It is far from obvious what is supposed to support this additional premise. Indeed, pending further explanation of what a 'direct' ground is, it is far from clear what the additional premise means.

In view of this unclarity, one might be tempted to delete the word 'direct' from the formulation of the Inversion Principle altogether, so that it now says simply: 'Whatever follows from any of the grounds for asserting a formula must follow from that formula.' The resulting position, however, does not at all fit the view we are considering, whereby the introduction rule for a connective is held to specify that connective's meaning. At least, it does not fit this view if the rules in question are to be the familiar introduction rules for the connectives. On this version of the view, the rule of $\vee$-introduction would imply that there are grounds for asserting the disjunction $\ulcorner A \vee B \urcorner$ if and only if there are grounds for asserting A or grounds for asserting B. And the 'only if' part of this claim is simply false, at least if the symbol '$\vee$' is supposed to have even roughly the same meaning as the English word 'or.' As Dummett noted in his early paper "Truth" (1959), the claim is wholly unsustainable if we allow that the testimony of others can provide grounds for assertions. Reliable sources from the Egyptian Fourth Dynasty tell us that the Pharaoh Cheops (whom Egyptologists now call 'Khufu') was either the son or the stepson of his predecessor on the throne, the Pharaoh Sneferu. Those sources, then, provide grounds for that disjunctive assertion. There are, though, no reliable grounds for asserting either disjunct. Indeed, there are other counter-examples to the 'only if' claim which do not rely on knowledge by testimony. If Inspector Morse knows (from the position of wounds on the victim) that the murderer is left-handed, and that Smith and Jones are the only left-handers among the possible culprits, then he has grounds for asserting 'Either Smith or Jones is the murderer.' In that circumstance, though, Morse may have no grounds for asserting either 'Smith is the murderer' or 'Jones is the murderer.' What we see, then, is that the present argument for the Inversion Principle depends upon finding a sense for the word 'direct' (or 'canonical') which treads a fine line. The sense has to be sufficiently generous to ensure that whenever a ground for asserting a formula obtains, a direct ground obtains. At the same time, it has to be sufficiently restricted to ensure that a direct ground for asserting $\ulcorner A \vee B \urcorner$ involves either a direct ground for A or a direct ground for B. The introduction rules for the other connectives will impose corresponding restrictions on the acceptable sense of 'direct.'

3 Problems with the Argument

What are the prospects of overcoming these problems so that the present argument for Inversion can be vindicated? I address the problems in turn.

There is no doubt that the conception of consequence on which the argument rests deviates from the conception which has animated logic since its creation. The key mark of a valid argument is that its conclusion is true whenever all its premises are true. At the heart of consequence, then, lies preservation of truth, not preservation of correct assertibility, or of knowability, or of anything other than truth. While disputes persist about the proper explanation of consequence, those disputes center on what surrounds that heart: notably whether consequence involves the necessary preservation of truth (as Aristotle held) or whether actual preservation will do (as Russell thought); and whether the sort of truth-preservation which is characteristic of logical consequence must be rooted in a formal relationship between premises and conclusion. If an explanation of consequence in terms of the preservation of correct assertibility is going to be more than an eccentric misuse of the familiar notion, we must take ourselves to be in a dialectical context in which truth has already been 'dethroned' (as people used to say) from its usual place in that explanation.

More particularly, we must presume that powerful arguments have already been given for explaining consequence in terms of the preservation of correct assertibility.

Supposing for a moment that some powerful arguments to this effect have been given, the foundations of logic will certainly need reconstruction. Logicians prove soundness theorems for various logical systems. That is, they show that, if the rules of a given system are followed, then the conclusion is true in every possible circumstance in which all the premises are true. But what can soundness come to if truth has been dethroned? There must still be some standard against which individual deductions, rules, and indeed whole logical systems may be assessed. We still want to be able to say that someone who infers 'If Fred works hard, he will get a First; Fred will get a First; therefore Fred works hard' has reasoned unsoundly – that he has made a logical mistake. But in what does his mistake consist if not in the possibility that both premises might be true when the conclusion is not true?

It may be answered that we can still give an account of why the reasoner is making a mistake in terms of correct assertibility. Our reasoner's argument is unsound because someone could be in a position correctly to assert both the premises without being in a position correctly to assert the conclusion. But this just pushes the problem back: we shall then need to specify the conditions for correctly asserting the sentences or formulae of the relevant language. On any view, the introduction of logical connectives into a language that has hitherto lacked them is going to create new grounds for asserting formulae. This applies to atomic formulae as well as to molecules: once the language contains a conditional, for example, we can correctly assert an atomic formula B by (for example) deducing it from correct assertions of A and of ⌜If A then B⌝. But given that any logical rules are going to generate new grounds for assertions, we have to say what it is for *modus ponens* to constitute an acceptable expansion of those grounds while affirming the consequent does not. Moreover, the proponent of the present argument for Harmony has to give an account of this matter without falling back on the idea that a valid argument preserves truth.

The only developed account of this that I know relies heavily on the distinction between direct and indirect grounds for assertion. The thought is that, while logic certainly yields new indirect grounds for atomic assertions, its rules must be faithful to the *direct* grounds of formulae: we shall have an instance of consequence only if any direct grounds for the premises could be transformed into a direct ground for the conclusion. This is the account of logical consequence shared by Prawitz (see especially his 1974) and Dummett (see especially his 1991). Instead of direct grounds for atomic formulae, Prawitz writes of valid 'closed' arguments for them. He duly "defines a sentence B as a logical consequence of sentences A_1, …., A_n by the existence of an operation φ which for every choice C [of valid closed arguments] transforms any closed arguments for sentences A_1, …., A_n valid relative to C to a closed argument for B valid relative to C" (Prawitz, 1974, pp. 74–75). Dummett proposes essentially the same account. "We regard [Euler's] proof as showing us, of someone observed to cross every bridge at Königsberg, that he crossed at least one bridge twice, *by the criteria we already possessed for crossing a bridge twice*" (1991, p. 219, emphasis in the original). "If that is what deductive inference achieves," he continues, "the requirement of harmony springs from its very nature. When an expression, including a logical constant, is introduced into the language, the rules for its use should determine its meaning, but its introduction should not be allowed to affect the meanings of sentences already in the language" (1991, p. 220). By mastering logic we acquire new indirect grounds for making assertions. But the methods we master must be faithful to the meanings of the atoms in that they preserve their conditions of direct assertibility.

If consequence is to be explained in terms of the preservation of some form of correct assertibility, it is hard to think of any other account than the one which Prawitz and Dummett provide. That account, though, generates serious problems – problems which, I now argue, are so serious as to cast doubt upon the hypothesis that consequence *can* be explained in this way.

Euler's proof is said to show us, of someone observed to cross every bridge at Königsberg, that he crossed at least one bridge twice, *by the criteria we already possessed for crossing a bridge twice*. But that cannot mean that those criteria have actually been applied to verify that our promenader crossed a bridge twice. Perhaps they were so applied – perhaps an observer stationed on the Dombrücke, for example, saw him cross that bridge twice – but Euler's proof would not be refuted if the pre-existing criteria were not actually applied. The most that can be claimed is a counterfactual: had an observer been stationed on each bridge, with instructions to tick a box if, and only if, the promenader was observed crossing it twice, then at least one observer would have ticked his box.

This counterfactual claim, however, is susceptible to objections parallel to those which afflict putative counterfactual analyses of other categorical notions. Some philosophers used to say that an object is yellow if an observer with good eyesight, viewing it in white light, would perceive it as yellow. Saul Kripke objected that this account was inconsistent with something that is surely a metaphysical possibility – namely, the existence of *killer yellow*, a shade of yellow that kills any observer who looks at it in white light.[9] In much the same way, Dummett's account of the validity of Euler's proof is inconsistent with the existence of *Königsberg ennui*, a strange neurological condition which ensures that anyone trying to observe whether a promenader has crossed a given bridge twice will fall into a catatonic state before any second crossing. Like killer yellow, Königsberg ennui is surely a metaphysical possibility. In a possible world where the denizens of Königsberg are afflicted by it, however, it will not be true to say that at least one of our observers would have ticked his box, had the promenader crossed every bridge. Even in such a world, however, 'The promenader crossed at least one bridge twice' still follows from 'He crossed every bridge.'

Even if we prescind from this rather general doubt, other worries press in fast, especially when we turn to our second main problem and reflect on the role which the distinction between direct and indirect grounds needs to play in the present argument for the Harmony Thesis. The notion of directness needs to be sufficiently generous, we said, that no ground for asserting a formula obtains unless a direct ground for asserting it could have obtained. Yet the direct grounds for asserting a complex formula are constrained to be those given by the introduction rule for the formula's main connective. Combining these points, we deduce that no ground for asserting a complex formula can obtain unless the assertion of that formula could have been justified by applying the introduction rule for its main connective. In other words, the present argument for the Harmony Thesis rests upon what Dummett calls the *Fundamental Assumption*.

Dummett is clear that the present argument does rest upon this Assumption. His discussion, though, does not inspire great confidence in the Assumption's truth. The Assumption is tenable, I think, in the case of conjunction. If someone is entitled to assert $\ulcorner A \wedge B \urcorner$, then he is entitled to assert A and is also entitled to assert B, so his assertion of the conjunction could have been grounded in an application of the $\wedge$-introduction rule. For none of the other familiar sentential connectives, though, is the Fundamental Assumption remotely plausible.

In the case of disjunction, Dummett recognizes that the Assumption is quite untenable if we confine ourselves to the grounds available to an individual thinker. While I have a

ground for asserting 'Cheops was either the son or the stepson of Sneferu,' it is impossible for me to justify that disjunctive assertion by an argument which concludes with an application of the introduction rule for 'or.' The Assumption is only tenable, Dummett holds, if the grounds for making an assertion are taken to include those available to any of *us*, where "whatever witnesses we trust must be included among 'ourselves'" (1991, p. 266). Thus the ancient scribe who recorded that Cheops was either the son or the stepson of Sneferu is one of us, and the Fundamental Assumption tells us that his assertion is correctly made only if he knew which it was, or if he was himself told the disjunction by someone who knew which disjunct obtained.

Perhaps we can swallow *this* consequence of the Assumption. Other consequences, though, are far less palatable. Consider the assertion 'At the moment when Brutus first stabbed Caesar in Pompey's Theatre, there was either an odd or an even number of people in the Agora in Athens.' Let us assume that the space of the Agora has been precisely delimited, and that precise rules have been laid down for when a person counts as being *in* a space. Given that assumption, most of would think ourselves entitled to make the present disjunctive assertion. If we are so entitled, though, the Fundamental Assumption entails that someone – that is, some one of us – could have been in a position either to assert 'At the moment when Brutus first stabbed Caesar, an odd number of people were in the Agora' or to assert 'At that moment, an even number of people were there.'

Dummett acknowledges, of course, that no one actually was in a position to make either of these claims. "To interpret the fundamental assumption," he writes, "we have to invoke the sense of 'could have' which was used earlier to characterize what may be called the minimal undeniable concession to realism demanded by the existence of deductive inference" (1991, p. 267). In the case of statements about the past, he continues, this means "that a sufficient condition for [an assertion's] correctness is that there exist effective means by which, at the relevant time, someone appropriately situated *could have* converted observations that were actually made into a verification of the statement asserted" (1991, p. 268). By the Fundamental Assumption, though, a closely related condition must be *necessary*: for an assertion to be correct, it is necessary that someone appropriately situated could, at the relevant time, have made observations which would have justified it. In the case of either of our two disjuncts, though, it is hard to see how this necessary condition could be satisfied. For where would an observer be 'appropriately situated'? An observer in Pompey's Theatre would have been well placed to notice when Brutus stabbed Caesar and to observe what was happening at that moment in that part of Rome; but he was not in a position to count how many people were then in the Athenian Agora. An observer situated in the Agora, by contrast, may have been in a position to make a count of those present in the square; but he would not know when to do so. What a direct ground for either disjunct needs is a *pair* of observers, with the first able to effect a practically instantaneous signal to tell the second when to make the count. But there was no effective means of sending such a signal 'at the relevant time': the necessary technology would not be invented for several centuries. Even if we gloss the Fundamental Assumption in the generous way that Dummett recommends, then, it is going to exclude many assertions that we take ourselves to be in a position to make. Its hard-line adherents may swallow that consequence. The rest of us, though, will simply conclude that the Fundamental Assumption is false when applied to disjunctions.

Matters are no better when we turn to (indicative) conditionals. Dummett himself concedes that he cannot defend the Assumption for conditionals with disjunctive consequents (see 1991, p. 273) but in fact the problem conditionals present for it runs far deeper: the

difficulty is that the standard introduction rule, Conditional Proof, is not a plausible codification of the circumstances in which we take ourselves to be entitled to assert English indicative conditionals. If Conditional Proof were the operative introduction rule for the vernacular 'if … then,' a direct ground for asserting a conditional would be a method for transforming any possible ground for the antecedent into a ground for the consequent, but this principle does not get the assertibility conditions of ordinary conditionals right. Variants of Moore's Paradox provide one class of counter-examples. Consider the conjunction 'It is raining but there are no grounds for asserting that it is raining.' It is plausible to hold that there are no possible grounds for asserting this conjunction: any grounds for asserting the first conjunct will falsify the second conjunct. Accordingly, we shall (vacuously) have a method for transforming any possible ground for this conjunction into a ground for asserting a self-evident absurdity, such as $0 = 1$. Given the principle about conditionals, it follows that there is a ground – indeed a direct ground – for asserting 'If it is raining but there are no grounds for asserting that it is raining, then $0 = 1$.' But that conditional does not seem to be one that we shall wish to assert: in entertaining the supposition or hypothesis 'It is raining but there are no grounds for asserting that it is raining' we do not seem to be entertaining an absurdity but something which might well be the case.

The crucial point here is that in a conditional we conditionalize on the *truth* of the antecedent, not on its assertibility. Ironically, in some of his other writings Dummett makes this point very clearly. "In a sentence like 'If you go into that room, you will die before nightfall,'" he remarks, "the event stated in the consequent is predicted on condition of the truth of the antecedent (construed as the future tense proper[10]), not of its justifiability" (1990, p. 193). As a point about the meaning of conditionals in English this is clearly correct, and Dummett goes so far as to conjecture that it is when statements occur as antecedents of conditionals (and in related complex constructions) that we need to draw the distinction between truth and justifiability (1990, p. 193). However that may be, the view that Conditional Proof specifies direct grounds for the assertion of ordinary conditionals is miles from the truth.

One might respond to this by saying that some *other* rule justifies such assertions; on the view we are considering, it will be this other rule which specifies the sense of the conditional. Even if it were possible to formulate an alternative rule, however, that would not help in the present context. For (a) we *do* seem to be prepared to eliminate vernacular conditionals using the rule of *modus ponens* while (b) it is *modus ponens* which stands as the inverse of Conditional Proof (for a proof see, e.g., Negri and von Plato, 2001, p. 8).[11] In other words, whatever exactly they are, the rules which we actually go by in introducing and eliminating vernacular conditionals are not in harmony.

Severe problems also afflict the Fundamental Assumption as it applies to negated statements. According to the Assumption, we shall not be entitled to assert a statement in the form ⌜Not A⌝ unless we could have justified that assertion by applying the introduction rule for 'not.' According to that introduction rule, we may assert ⌜Not A⌝ when we have derived a contradiction from our premises along with the hypothesis A. In many circumstances where we take ourselves to be entitled to assert ⌜Not A⌝, however, it is hard to see what the appropriate premises might be. I look out of the window and see that it is not raining. I am surely entitled to assert 'It is not raining,' but what premises does my observation deliver that would enable me to justify that assertion by applying the rule of 'not'-introduction? In many circumstances of this kind, there is no plausible answer. In looking out of the window, I might see that it is sunny, but being sunny is compatible with rain. The only specification of the content of my experience that is guaranteed to be incompatible

with 'It is raining' is 'It is not raining,' but while I can indeed see that it is not raining, the ensuing belief that it is not raining serves as a *premise* in my reasoning. It is not a *conclusion* that has been reached by applying the rule of 'not'-introduction to some other premises.

In fact, the situation with negation is even worse than that with disjunction and the conditional. In stating the introduction rule for negation, I said that ⌜Not A⌝ may be derived from some premises when the combination of those premises with A yields a contradiction. But what is a contradiction? One answer might be: it is any statement in the form ⌜A and not A⌝ – but we shall know that such a statement is contradictory only if we already know what 'not' means, so we cannot invoke this notion of a contradiction in a rule which purports to give the sense of 'not.' What is worse, if we understand the term 'introduction rule' in the way proposed in §1, it is demonstrable that there is no classically or intuitionistically correct introduction rule for '¬.' More generally, let us follow Humberstone and Makinson in calling a k-place connective **C** *contrarian* if $\mathbf{C}(P_1,\ldots,P_k)$ is valued *False* when all of $P_1,\ldots,P_k$ are valued *True*. (Thus the *falsum* '⊥,' conceived as a zero-place connective, and the unary connective '¬' are both contrarian in this sense.) Then there is no classically or intuitionistically correct introduction rule for any contrarian connective.[12] For let **C** be such a connective and suppose its introduction rule comprises all instances of the scheme

$$< X_1 \Rightarrow A_1,\ldots,X_n \Rightarrow A_n / Y \Rightarrow \mathbf{C}(P_1,\ldots,P_k) >.$$

Since the rule is an introduction rule, it is elementary, so all the formulae in the premise sequents $X_1 \Rightarrow A_1,\ldots, X_n \Rightarrow A_n$ and in the set Y must be sentence letters. But then the rule cannot be classically correct. Consider the substitution instance got by replacing each sentence letter by a classical tautology: under this substitution, each premise sequent becomes classically valid while the conclusion sequent has antecedents that are all true but a false succedent. Since every intuitionistically correct rule is also classically correct, there is no intuitionistically correct introduction rule for a contrarian connective either.

This result may seem bizarre: we teach our logic students sequent rules for '¬,' after all. On reflection, however, it is no surprise that sequent rules *of the form described* cannot characterize the logically relevant meaning of '¬.' Such rules ensure the correctness of certain sequents – that is, they ensure that *if* certain antecedents are true, then so are certain succedents. No collection of such rules can exclude the possibility that all the formulae in the language L are true, but we need to exclude that possibility in order to characterize '¬,' or indeed any contrarian connective. We need to ensure, for example, that ⌜$A \wedge \neg A$⌝ is *not* true.

The operative conception of an introduction rule needs to be liberalized, then, if a contrarian connective is to possess one. From a formal point of view, the simplest and most common liberalization permits a sequent to have a null succedent. We move, in other words, from *set/formula* sequents to *set/formula-or-empty* sequents. Such a sequent is correct if and only if the formula in its succedent is true whenever every formula in its antecedent is true. When the succedent is empty, that is, when there is no formula in it, the sequent will be correct if and only if not every formula in its antecedent is true. The sequent, $Q \Rightarrow \varnothing$, for example, will then be correct if and only if Q is not true. When the logical rules regulate set/formula-or-empty sequents, it is straightforward to give an introduction rule for '¬,' namely, $X \cup \{P_1\} \Rightarrow \varnothing / X \Rightarrow \neg P_1$.[13] Indeed, as Makinson (2014) points out, in such a system we can give an introduction rule for any truth-functional connective apart from '⊥.' We should not expect '⊥' to have an introduction rule.[14] On the theory we are considering, such a rule would specify the canonical conditions for asserting '⊥'; it would be surprising if there were conditions in which a speaker would be entitled to assert a formula which is

understood always to be false. For any connective **C** not equivalent to '⊥,' however, there will be at least one structure v where $v(\mathbf{C}(P_1,\ldots, P_k))$ is true. Where $P_{j_1},\ldots, P_{j_m}$ are those sentence letters evaluated as true under v and $P_{l_1},\ldots, P_{l_n}$ are those evaluated as false there, we have corresponding to v the rule

$$<\varnothing \Rightarrow P_{j_1},\ldots, \varnothing \Rightarrow P_{j_m}, P_{l_1} \Rightarrow \varnothing,\ldots, P_{l_n} \Rightarrow \varnothing / \varnothing \Rightarrow \mathbf{C}(P_1,\ldots, P_k)>.$$

The union of such rules for all v where $v(\mathbf{C}(P_1,\ldots, P_k))$ is true is then the introduction rule for **C**.[15]

Natural as this liberalization is from a formal point of view, it comes at a philosophical price. As remarked at the outset, many adherents of the Harmony Thesis are also adherents of Inferential Role Semantics. As such, they are ambitious to characterize an expression's meaning by the rules that regulate its inferential *use*. The move from set/formula sequents to set/formula-or-empty sequents, however, involves a retreat from direct engagement with the way logical expressions are used in inference. A set/formula sequent represents an actual argument, in which a reasoner passes from a set of premises to a conclusion. Hence the correctness of such a sequent can be related to the intuitive acceptability of the corresponding inferential passage. Where a speaker fails to reach a conclusion, however, we do not have an inference; we merely have a list of statements. Accordingly, we cannot explain the correctness of a set/formula-or-empty sequent directly in terms of the intuitive acceptability of an inference. We shall need instead to give a metalogical account of correctness, such as that in the previous paragraph. This takes us further away from what, for an IRS theorist, is foundational.

There are, to be sure, alternative ways of liberalizing the formal system which cleave more closely to the ideal that its rules should record the way we use connectives. I expounded one of these in my essay '"Yes" and "No."'[16] The operational logical rules given there are 'bilateral' principles which regulate deductive transitions between premises and a conclusion in each of which a yes-no question is followed by one of its expected answers, as in 'Is Fred in Berlin? No. So is it the case that he is either in Paris or is not in Berlin? Yes.' But even if we find a way of remaining faithful to this ideal, the present strategy for justifying the Harmony Thesis has reached a dead end. Dummett conceded that his "examination of the fundamental assumption has left it very shaky" (1991, p. 277), and with this conclusion we can only concur. A theory of the meaning of the connectives which passes muster for 'and,' but which fails for 'or,' 'if … then,' and 'not' – which is committed, indeed, to counting these ubiquitous expressions as meaningless – is not doing at all well.

One might wonder how Dummett felt entitled to pursue his project of justifying the laws of intuitionistic logic, while delegitimating classical logic on the ground of its alleged violations of the Harmony Thesis, when he admitted that the grounds for the Thesis were so shaky. His reason is interesting. The laws of intuitionistic logic, he says, "are not going to be called into question by any uncertainties over the scope or status of the fundamental assumption, precisely because the classical logician will admit that assumption, interpreted in terms of an ideal observer" (1991, p. 279). His thought seems to have been this. At this point in the dialectic, we shall have been so completely persuaded – perhaps by Dummett's own 'manifestation argument,' or his argument about the acquisition of language – that truth needs to be dethroned from its place in the traditional explanation of consequence, that we shall accept the Prawitz–Dummett account of that notion in terms of preservation of direct grounds. It is simply that those people willing to assert (for example) the disjunction 'At the moment when Brutus first stabbed Caesar, there was either an odd or an even number of people in the Agora'

will do so because they are prepared to postulate a God-like ideal observer who was able to make a count of the people in the Agora in Athens at the very moment when the knife went in in Rome. The upshot of our discussion, however, is that this is quite the wrong moral to draw. There is no space to rehearse Dummett's arguments against the intelligibility of a notion of truth that goes beyond the existence of a verification. To put the point at its mildest, though, those arguments are far from conclusive, and our analysis suggests that a real strike against their conclusion is the immense difficulty we shall then face in trying to forge a notion of consequence to replace the familiar one cast in terms of truth-preservation.

More precisely, what we have seen are some of the difficulties in forging an alternative account of consequence that will sustain the Harmony Thesis. At every turn, the traditional account in terms of preservation of truth cries out to be restored. Of course, there is no suggestion that resurrecting the traditional account is going to open any direct path towards justifying classical logic: for one thing, one can adopt the traditional account of consequence without postulating the Principle of Bivalence. But the arguments we have so far considered for imposing a harmony requirement, and for deviating from classical logic because it violates that requirement, lead only into a morass. If this is the best that can be said in favor of the Harmony Thesis, the classical logician has nothing to fear from it.

4 Arguments from the 'Innocence' of Logic

The argument for the Harmony Thesis just analyzed assumes that the introduction rule for a connective specifies its meaning. That assumption is in any case far from compelling. It is more plausible to take ordinary competence with the indicative conditional, for example, to be manifest in applications of its elimination rule – *modus ponens* – than in mastery of whatever rule regulates its introduction.[17] The arguments for Harmony that I consider next do not assume any semantic priority for the introduction rules. Indeed, they do not assume that any one sort of rule – whether it be the introduction rules or the elimination rules – will by itself specify the meanings of the connectives.[18]

The first such argument rests on a premise about the nature of logic. Dummett sometimes writes as though the Inversion Principle follows from a general requirement of harmony that applies between the grounds and consequences of any meaningful statement. Florian Steinberger, *per contra*, argues as follows:

> Whatever misgivings one may have about Dummett's wider project, a strong case can be made for a logic-specific harmony requirement. The reason for this stems from the role logic plays in our assertoric practices. On the use-theoretic view [of meaning], the meanings of non-logical sentences (sentences not containing any logical operators) are thought to be given by their I- and E-principles.[19] Logic, in addition to the direct grounds for assertion given by the appropriate I-principles, offers indirect grounds for asserting non-logical sentences: we may assert a non-logical sentence if it can be correctly deduced from a set of accepted premisses. But for these indirect deductive routes to assertibility to be not only legitimate but to have the unassailable reliability we require of logical inference, our logical modes of inference must respect the conditions under which the (direct) assertion of non-logical sentences is justified. That is, logical inference alone may not license the assertion of non-logical sentences that we should not have been in a position to assert directly (at least in principle). Let us call this the *principle of innocence*: it should not be possible, solely by engaging in deductive logical reasoning, to discover hitherto unknown (atomic) truths that we would have been incapable of

> discovering independently of logic…. How can we make sure that innocence obtains? This is where harmony comes in. The primary purpose of harmony is to secure the innocence of logic. (Steinberger, 2011, pp. 619–620)

Steinberger further contends that harmony is the best way – perhaps the only way – of securing innocence:

> A moment's reflection reveals not only that harmony is an adequate measure, but that it seems entirely natural that any measure designed to guarantee the holding of the requirement of innocence should take the form of a harmony requirement. After all, our aim is to ensure that the meanings of the logical constants are fixed in such a way as not to perturb the non-logical regions of language. The best way to do this (at a local level) is by requiring that the introduction and elimination rules that govern the meanings of logical constants be exactly commensurate in strength. Why? Well, because when such an equilibrium between I-rules and E-rules obtains, we can rest assured that our deductive practices will not, as it were, create novel grounds for asserting non-logical sentences (as in the case, for example, of [Prior's invented connective] *tonk*). The requirement of harmony thus seems to be an eminently reasonable and natural safeguard for the principle of innocence. (Steinberger, 2011, p. 620)

Steinberger spells out as follows the equilibrium he has in mind: "nothing more and nothing less may be deduced from an assertion of $A\$B$ via \$-E than can already be deduced from the premisses of the corresponding I-rules. Put another way, … the E-rules ought to exploit *all* and *only* the inferential powers that the I-rules have bestowed upon it" (Steinberger, 2011, p. 620). His requirement of equilibrium, then, demands the satisfaction of Negri and von Plato's Inversion Principle, and something more besides. As we saw in §1, their Inversion Principle is satisfied when the standard introduction rule for '$\vee$' is paired with its quantum-logical elimination rule, but Steinberger insists that that elimination rule fails to exploit *all* the inferential power bestowed by the introduction rule (see his discussion of 'E-weak disharmony,' 2011, p. 621). If it works, then, Steinberger's argument justifies a form of the Harmony Thesis that is even more exacting (as far as the logical connectives are concerned) than that proposed by Prawitz and Dummett.

Does his argument work, though? There are strong reasons to doubt it.

First, the principle of innocence is far less compelling than Steinberger supposes. Since Mill's *A System of Logic*, with its notorious claim that "nothing ever was, or can be, proved by syllogism which was not known, or assumed to be known, before" (Mill, 1891, II iii 1), a central problem in the philosophy of logic has been to reconcile the conclusiveness of correct deduction with its ability to expand our knowledge. Part of the explanation of how deduction generates new knowledge is that premises founded on different sources of knowledge can entail a conclusion that no source founds by itself. A trusty informant tells me that John is either in the common room or in the library; I see that he is not in the common room; I deduce that he is in the library. I know the first premise through testimony and I know the second as result of observation. But whilst I come to know the conclusion, my knowledge of it does not stem from either testimony or observation alone: I was not told that John is in the library, nor did I see him there. Of course, this case is not itself a counter-example to Steinberger's principle of innocence. Although I did not in fact see John in the library, in principle I could have done. In our ordinary deductive practice,

however, we are fully prepared to splice together different sources of knowledge to deduce conclusions that could not be founded on any 'direct' evidence, even in principle. Suppose I know – through astronomical theory, and appropriate observations – that a body *B* is either in region *R* of the Andromeda Galaxy or is in a black hole. Suppose I make some further observations, and come to know that *B* is not in region *R*. I may then deduce that *B* is in a black hole. It is (we may assume) impossible even in principle to discover by direct observation whether a body is in a black hole. For all that, this second deduction appears to be just as cogent as the first. In each case, the deduction yields knowledge of its conclusion, even though in the second case such knowledge could not have been attained directly, even in principle. *Contra* Steinberger's principle of innocence, then, by engaging in deduction we can discover hitherto unknown atomic truths which we would have been incapable of discovering without logic. The principle of innocence is far from innocent. Were it accepted, it would seriously restrict the use we actually make of deductive logic in enlarging our knowledge.

Second, even if we grant the principle, it may be secured by a weaker requirement than that of equilibrium between introduction and elimination rules. As Steinberger acknowledges, innocence will be secured if the non-logical regions of language are unperturbed by the logical rules – that is, if those rules create no novel grounds for asserting non-logical (i.e., atomic) sentences. This condition will be met if the logical rules *en bloc* create no such novel grounds – a 'global' condition in his terms. To ensure innocence, then, we need not descend to the 'local' level and require the introduction and elimination rules of each individual connective to be in equilibrium.

There is, in fact, a natural way of making the global condition more precise. The direct grounds for asserting non-logical formulae will induce a consequence relation R^- on the sublanguage L^- that comprises only such formulae. Innocence will be secured if the expanded consequence relation *R* that is induced on the full language *L* when the logical rules are added is a *conservative extension* of R^-. That is, *R* as restricted to L^- does not extend R^-. Steinberger is well aware that this global condition may be satisfied even when the introduction and elimination rules of a certain connective are not in equilibrium (2011, pp. 625 and 634–638). The problem is that his principle of innocence only sustains the global condition. The conservative extension requirement is enough to ensure that the 'meanings' of atomic formulae – or better: the consequential relations between them – are left unperturbed. It also excludes Prior's rogue connective 'tonk,' whose introduction rule is $< P_1 / P_1$ tonk $P_2 >$ and whose elimination rule is $< P_1$ tonk $P_2 / P_2 >$ (Prior, 1960). 'Tonk' is indeed a runabout inference ticket, which licenses the move from one formula to any other, so its rules will violate conservativeness unless the pre-logical consequence relation R^- is already total (see Belnap, 1962).

Robert Brandom's conception of logic is similar to Steinberger's, but he is more circumspect about its implications for harmony (Brandom, 1994; 2000). On Brandom's view, what characterizes the logical notions is their role in 'making explicit' the relations between non-logical sentences that are traced out in material inferences and in recognitions of incompatibility. A good material inference takes a thinker from 'Edinburgh is north of London' to 'London is south of Edinburgh'; we may express our acceptance of that inference by asserting the conditional 'If Edinburgh is north of London, then London is south of Edinburgh.' Similarly, we express our recognition of the incompatibility between the table's being red all over and its being green all over by saying, 'If the table is red all over,

then it is not green all over.' What Brandom takes to follow from this doctrine is, simply, the conservative extension requirement:

> Unless the introduction and elimination rules are inferentially conservative, the introduction of the new vocabulary licenses new material inferences, and so alters the contents associated with the old vocabulary. So if logical vocabulary is to play its distinctive expressive role of making explicit the original material inferences, and so conceptual contents expressed by the old vocabulary, it must be a criterion of adequacy for introducing logical vocabulary that no new inferences involving the old vocabulary be made appropriate thereby. (Brandom, 2000, pp. 68–69; see Brandom, 1994, pp. 123–130 for a fuller exposition of the same argument)

Brandom is right, I think, to claim that his expressivist account of logic demands the conservative extension requirement. The important point for present purposes, however, is that it *only* demands that: conservative extension may be satisfied even if the rules for the connectives do not satisfy the Principle of Inversion, that is, even if the Harmony Thesis is false.[20]

This comes out clearly, indeed, in the case of the classical propositional calculus. As noted earlier, the introduction rule for negation is *Simple Reductio*: $X \cup \{P_1\} \Rightarrow \varnothing / X \Rightarrow \neg P_1$. The elimination rule which is in harmony with this is *Ex Contradictione Quodlibet*: $\{P_1, \neg P_1\} \Rightarrow P_2$. While these two rules together characterize the intuitionistic logic of negation, its classical logic demands a further principle. There are many additional principles which will do. For definiteness, let us add the rule form of Excluded Middle,

$$EM: \quad X \cup \{A\} \Rightarrow B,\ Y \cup \{\neg A\} \Rightarrow B / X \cup Y \Rightarrow B.$$

(Adding assumptions to a sequent is often called 'thinning,' so Harold Hodes aptly calls *EM* a 'thickening' rule: it allows the deduction of a sequent from other sequents with the same succedent whose antecedents include formulae which do not appear in the antecedent of the conclusion (Hodes, 2004, p. 148)). In the resulting system, the Harmony Thesis is violated: the logical behavior of '¬' is not regulated *only* by a pair of harmonious introduction and elimination rules. But the classical consequence relation conservatively extends whatever pre-logical consequence relation obtains among the non-logical formulae: where $P_1, \ldots, P_n$ and Q are atoms, there is a classically admissible truth-value assignment which assigns *True* to all of $P_1, \ldots, P_n$ and False to Q, so if Q is a consequence of $P_1, \ldots, P_n$, that must be because the pre-logical consequence relation determines it as such.[21]

5 Tennant's Argument for Harmony

Neil Tennant is an adherent of the Harmony Thesis who distinguishes sharply between harmony proper and the conservative extension requirement.[22] He also holds that the introduction rule and elimination rule for a connective *jointly* determine its sense, so that the Thesis cannot be justified by requiring the rules of one sort to keep faith with the meanings already laid down by rules of the other sort.[23] Rather, he holds, its justification arises from a requirement of coherence between the introduction and elimination rules for a given connective. As Prior's 'tonk' shows, not every pair of introduction and elimination rules succeeds in endowing the connective it purports to characterize with a sense. It is by spelling out the requirements for coherence that Tennant aims to justify the Thesis.[24]

The precise form of harmony that Tennant defends is subtly different from that in Negri and von Plato. To state it accurately we need some terminology. Let us say that a formula is a *maximally strong F* if it is *F* and entails any formula that is *F*; and let us say that a formula is a *maximally weak F* if it is *F* and is entailed by any formula that is *F*. Tennant's First Principle of Harmony may then be stated as follows. A formula whose main connective is **C** is to be

(a) a maximally strong statement that can stand as conclusion of the introduction rule for **C**, given the elimination rule for **C**; and
(b) a maximally weak statement that can stand as major premise of the elimination rule for **C**, given the introduction rule for **C**.

We may illustrate this First Principle with the case of '$\vee$.' To show part (a) for this connective, let *X* be any formula that can stand as the conclusion of $\vee$-introduction, with *A* and *B* as premises. We then have that *A* entails *X* and *B* entails *X*. By $\vee$-elimination, it follows that $\ulcorner A \vee B \urcorner$ entails *X*, showing that $\ulcorner A \vee B \urcorner$ is a maximally strong statement that can stand as conclusion of $\vee$-introduction. To show part (b), let *Y* be any formula that can stand as the major premise of $\vee$-elimination. Then, whenever *A* entails *C* and *B* entails *C*, *Y* entails *C*. By $\vee$-introduction, *A* entails $\ulcorner A \vee B \urcorner$, as does *B*, so *Y* entails $\ulcorner A \vee B \urcorner$. Thus $\ulcorner A \vee B \urcorner$ is a maximally weak statement that can stand as major premise of $\vee$-elimination.

As this demonstration shows, Tennant's First Principle is satisfied whether the $\vee$-elimination rule has its usual form, in which the use of side premises is permitted, or takes the restricted form it has in quantum logic, in which *C* may be inferred from $\ulcorner A \vee B \urcorner$ only if it follows from *A* alone and from *B* alone. The First Principle, then, fails to distinguish between the two forms of the rule. For this reason, Tennant also lays down a Second Principle. When a pair of introduction and elimination rules **CI** and **CE** for a connective **C** meets conditions (a) and (b), let us say that the pair is in harmony (with a small 'h'). We further say that the pair is in Harmony (with a capital 'H') if **CE** is the strongest elimination rule with which **CI** is in harmony and **CI** is the strongest introduction rule with which **CE** is in harmony. Tennant's Second Principle requires the rules for a meaningful connective to be in Harmony. The form of $\vee$-elimination which permits side premises is stronger than the form which does not: it allows us to derive more conclusions from a given disjunction. It is the unrestricted form of the elimination rule, then, which is in Harmony with the rule of $\vee$-introduction, so the restricted form falls foul of Tennant's Second Principle. It is good to have a criterion which improves on Negri and von Plato's Inversion Principle in excluding the restricted form of $\vee$-elimination. We shall, however, need a justification for requiring rules to be in Harmony as well as harmony.

Tennant contends that satisfaction of both his Principles of Harmony is 'a *conditio sine qua non* of' rules specifying or constituting a coherent meaning for the connective in question (1987, p. 94). He further claims that this condition has revisionary implications for logic. "The correct consequence relation, insofar as it should arise solely from the meanings of the logical constants, is, naturally, the least relation with respect to which the Harmony of the rules governing those constants can be sustained" (1987, p. 97). The least such relation, Tennant thinks, is that characterized by the system he calls intuitionistic relevant logic. "I intend thereby to reveal as unjustifiable excrescences the extra ingredients in the consequence relation of classical logic that have earned the generic labels (a) the fallacies of relevance, and (b) the classical laws of negation" (1987, p. 97). We need, then, to consider the arguments Tennant advances for two theses. The first thesis is his claim that satisfaction of the two Principles of Harmony is a necessary condition for a connective to possess a coherent meaning.

The second is the claim that any logical principles that go beyond a Harmonious pair of introduction and elimination rules are 'unjustifiable excrescences.' I shall contend that neither of these theses is well supported.[25]

How does Tennant argue for the first thesis? The precise course of his reasoning is somewhat hard to follow, but we are told that "the requirement for harmony emerges clearly if one follows a philosophical method that has the appearance of empirical speculation about the origins of language, but is actually designed to focus on constitutive features of meaning. This is the method of enquiring after the *aetiology of entrenchment* of expressions in a language and of conventions governing their use" (1987, p. 77). The general idea is that we shall be unable to explain how the logical connectives could have become entrenched in a language – that is to say, how they could have acquired a stable meaning – unless their introduction and elimination rules are in Harmony. The claim that Harmony is a *conditio sine qua non* for such rules to constitute or specify a connective's meaning duly comes at the end of a long passage describing how meanings for the connectives might have become entrenched.

What is Tennant's account of entrenchment? It certainly has the appearance of empirical speculation about the origins of language. We are asked to imagine a community of speakers who start off using only atomic statements; Tennant then asks how connectives could be added – one at a time – to their dialect. He suggests that we would be unable to understand how this could happen unless the rules governing those connectives satisfy his two Principles of Harmony.

This account exemplifies a genre which one might call the *Just So Story*. We find it hard to imagine how a meaningful connective could have been added to a language unless certain conditions are met. So we take those conditions to be necessary for the connectives to have a meaning. Of course, my name for the genre carries a warning. Kipling's account of how the elephant got its trunk does have a certain explanatory charm. Few people today, though, would regard it as even a remote approximation to the truth. So if we are, O Best Beloved, to venture forth to the philosophical tributary of the great grey-green, greasy Limpopo River, all set about with normalized proof trees, we shall need to take care. We shall need to make sure that if a condition is imposed on the connectives, the condition really is necessary for them to have a meaning and does not simply express a philosopher's preconceptions about how language ought to work. If we read Tennant's account with that warning in mind, we shall find his story even less persuasive than Kipling's.

The root of the problem is that Tennant's account of how the connectives get entrenched does not explain how they come to have their actual meanings. The difficulty comes out clearly in the case of negation – a case which is of course central to the choice between classical and intuitionistic logic. Our signs for negation, Tennant hypothesizes, originate in the need one speaker may have to contradict or challenge an atomic assertion by another: "dialogue, not monologue is where negation first flourishes" (1987, p. 83). Let us accept this for the sake of argument. "The challenger," he goes on, "must have information to the contrary, rather than be merely playing the role of the *uninformed* doubter" (1987, p. 84). Let us accept this too. Tennant further contends that a speaker who challenges A by saying $\ulcorner$ Not A $\urcorner$ is "saying something *about the same subject matter* as A" (1987, p. 84, emphasis in the original). If a speaker who says $\ulcorner$ Not A $\urcorner$ were merely doubting "the existence of a warrant for [A], then the challenge would be self-warranting, for nothing could serve as better evidence for such a claim than its own making" (1987, p. 84). Again, this seems right. Tennant infers from this that the sort of challenge to A that is expressed by $\ulcorner$ Not A $\urcorner$ "must be conceived of as possessing warrants that are as open to independent public assessment

as are the warrants of the assertions challenged" (1987, p. 84). We may accept this too. "Denial of *A*," he concludes, "has the force 'I have good reason to believe that there is no warrant for *A*' rather than the weaker 'I have no reason to believe (apart from your asserting it) that you have any warrant for *A*.' Denial … carries with it no in-built guarantee of excluded middle" (1987, p. 85).

That final conclusion, though, does not follow from the considerations that are adduced to support it, and in any case Tennant seriously misrepresents the way ordinary speakers actually use signs for negation. Of course Tennant is right to claim that someone who says ⌜ Not *A* ⌝ is saying more than 'I have no reason to believe that you have any warrant for *A*.' As Heyting pointed out long ago, if this were the right account of the meaning of 'not,' then someone who said 'Not every even number greater than two is the sum of two primes' would be making an autobiographical statement, not a mathematical one. But Tennant's account equally misrepresents the content of that negated claim. On his view, someone who makes the claim is saying 'I have good reason to believe that there is no warrant for the claim that every even number greater than two is the sum of two primes.' Now in certain circumstances that might be a perfectly sensible thing to say, and if one understands negation in this way, then the law of excluded middle will indeed fail to be valid. It is not, however, the way most of us understand negation. On Tennant's account, it would be correct to say 'Not every even number greater than two is the sum of two primes' if the Goldbach Conjecture were unprovable. For most of us, though, it would be correct to say as much only if the Conjecture were false – that is, only if some even number greater than two were not the sum of two primes.

It may be replied that to object in this way is to fail to take seriously the possibility that classical logic might need to be revised. Not so: the objection is simply to Tennant's argument for revising it. It is, we are told, impossible to understand how the use of 'not' could have become entrenched unless it originated in the way Tennant describes. But the story he tells fails to explain the patterns of use which have actually become entrenched. In this respect, his Just So Story is less persuasive than Kipling's, for Kipling was at least offering an explanation for something that is actually the case. Elephants, after all, really do have trunks.

So much for Tennant's first thesis. What about his claim that any logical principles that go beyond a Harmonious pair of introduction and elimination rules are 'unjustifiable excrescences'? As far as I can discern, the only argument he gives for this second thesis is in a parenthesis: "The correct consequence relation, insofar as it should arise solely from the meanings of the logical constants, is, naturally, the least relation with respect to which the Harmony of the rules governing those constants can be sustained" (1987, p. 97). That 'insofar as' clause is doing all the work. Tennant seems to take it to be obvious that the logical laws regulating a connective will arise solely from its meaning, but he gives no argument for this claim, which is in truth very far from obvious. Certainly, it is not obvious that only introduction and elimination rules may regulate a connective's logical behavior. A classical logician might hold the position sketched at the end of §4, whereby the introduction and elimination rules for negation are the intuitionistic ones – namely, the rules of *Simple Reductio* and *Ex Contradictione Quodlibet* – but where an additional rule concerning negation – the thickening rule *EM* – is nevertheless valid.[26] There is no question of trying to justify *EM* by way of harmony considerations: it has the form neither of an introduction rule nor of an elimination rule. However, in the absence of any argument for the claim that the only valid principles that concern a connective are a Harmonious pair of such rules, there is no basis for Tennant's claim that *EM* is an 'unjustifiable excrescence.'

Tennant may protest that the case of 'tonk' shows that there must be some constraints on the introduction and elimination rules for a meaningful connective. Many people believe that those constraints amount to Harmony. Adding extra logical rules for a connective threatens to destabilize the equilibrium that Harmony guarantees, and thereby deprive the connective in question of a coherent sense. But there is a far better explanation of why 'tonk' is meaningless than that it violates Harmony. It is not meaningful, because a formula whose main connective it is does not *say* anything; such a formula does not say anything because it does not have truth-conditions. Thus consider the formula '2 is prime tonk 4 is prime.' This formula follows by 'tonk'-introduction from '2 is prime,' which is true, so it must be itself true. Yet '4 is prime,' which is not true, follows by 'tonk'-elimination from it, so the formula cannot be true. No coherent truth-condition can be assigned, then, to '2 is prime tonk 4 is prime,' and since both components do have truth-conditions, the culprit is clearly 'tonk.' As an explanation of why 'tonk' is meaningless, this explanation is superior to Belnap's, according to which 'tonk' is defective because it non-conservatively extends the pre-existing consequence relation. We shall sometimes want to do that, as when we add a truth-predicate to a mathematical theory (see n. 21 above), but we shall never want a declarative formula to lack truth-conditions. If a formula succeeds in saying something, it will have a truth-condition, *viz*., the condition that is satisfied if, and only if, things are as the formula says they are. So the only declarative formulae that lack truth-conditions are those that fail to say anything.[27]

6 Harmony and Inferential Role Semantics

Our examination of three prominent arguments for the Harmony Thesis has left it without any justification. Supposing it is false, does that threaten IRS? I think not. Our discussion of the Dummett–Prawitz argument for Harmony revealed the huge difficulties that confront the project of trying to explicate the notions of consequence and validity *directly* in terms of the rules which, for the IRS theorist, constitute the meanings of the connectives. But the IRS theorist is free to take an indirect approach. He might take the rules that characterize a connective's inferential role as specifying its sense, but allow that it also has a reference, or a semantic value. This semantic value will be the contribution the connective makes to the truth-conditions of a formula in which it occurs. Once we have a specification of truth-conditions for formulae of the relevant language, we can apply the traditional account of consequence in terms of the preservation (or necessary preservation) of truth.[28]

This oblique approach plainly requires an account of *how* the inferential roles determine semantic values – in Fregean terms, how sense determines reference. That is, it requires what Christopher Peacocke calls a 'determination theory.' In his paper "Understanding logical constants: a realist's account" (1987), Peacocke begins to develop such a theory for the connectives, and Hodes (2004) has pursued the matter further. The best hope of an IRS theory of meaning lies, I think, with this approach, and the determination theory goes more smoothly if the inferential roles played by the connectives are characterized by the 'bilateral' rules mentioned at n. 16 above, rather than the more familiar 'unilateral' rules. The kernel of any determination theory for the connectives will be the principle that the rules the reasoner goes by (or ought to go by) must preserve the correctness of sequents. For unilateral sequents, correctness is in turn a matter of preserving truth. Even given Bivalence, however, this constraint on the classical sequent rules fails to ensure that ⌜Not A⌝ is true whenever

A is false (see Peacocke, 1987, p. 164, and Hodes, 2004, p. 162). By contrast, that fact about the semantic value of 'not' may be 'read off' the intuitive correctness of the bilateral sequent rule exemplified by 'Is Fred at home? No. So is it the case that Fred is not at home? Yes.'

Whether a fully satisfactory determination theory can be given for the connectives is an open question – one of the most interesting and pressing in the philosophy of logic and language. The verdict on the immediate issue, though, is clear. Some people like Górecki's Third Symphony but few would say that it is a patch on Beethoven's. One reason is that Beethoven knew better when to leaven harmony with dissonance. As in music, so in logic: there is no universal requirement of harmony.

Notes

1 This chapter derives from an address delivered to a conference on Logic and Inference held at the University of London on March 20, 2015. I am much indebted to the organizers, Julien Murzi and Florian Steinberger, and to those who participated in the discussion: Sinan Dogramaci, Gilbert Harman, Gail Leckie, David Makinson, and Joshua Schechter. I also thank the editors of this volume for helpful comments on a draft.

2 On my account, then, introduction and elimination rules are rules in a sequent calculus. Some of the writings to be examined below take harmony to be a relation between rules in natural-deduction formalizations of logic. In a rigorous treatment of this topic, however, it is best to work in a sequent framework, where the assumptions on which a conclusion depends are explicitly represented. The philosophical arguments for harmony proposed by those who prefer natural-deduction formalizations transpose to the sequent framework.

3 For these definitions, cf. Humberstone and Makinson (2012, §2). As they remark (§2, n. 5), a rule which is elementary in the present sense will be both 'pure' and 'simple' in the terminology of Dummett (1991).

4 In the first edition of *Elements of Intuitionism* (1977, pp. 394–395), Dummett argued that the theory could be made compositional, all the same. For skepticism about his proposed way of achieving this, see Prawitz (1987, esp. pp. 156–163) and Pagin (2009, esp. pp. 724–734). Dummett entirely rewrote this passage for the second edition of *Elements*, and there concluded that the form of compositionality that could be justified was only "a very thin one" (2000, p. 274).

5 On the history of inversion principles, with references to Lorenzen (1950; 1955) and Schroeder-Heister (1984) as well as to Gentzen (1935) and Prawitz (1965), see Moriconi and Tesconi (2008).

6 Thereby gratifying a *desideratum* of Gentzen's: "By making these ideas more precise it should be possible to display the *E*-inferences [i.e., the elimination rules] as unique functions [*eindeutige Funktionen*] of their corresponding *I*-inferences [introduction rules], on the basis of certain requirements" (Gentzen, 1935, p. 189/1969, p. 81).

7 The same problem attends Stephen Read's requirement of 'general-elimination harmony.' See Read (2010, p. 566).

8 As Negri and von Plato recognize, their Inversion Principle yields more general forms of $\wedge$-elimination and of $\rightarrow$-elimination than one usually finds in the textbooks (see 2001, pp. 6–7 and 8–9). I do not object to this aspect of their theory, which might well be a bonus rather than a drawback. However, the inability to justify the unrestricted form of $\vee$-elimination is a difficulty.

9 Kripke presented this case in lectures which remain unpublished, but Lewis (1997) contains a brief account of it. Kripke has long been on record as an opponent of counterfactual and dispositional accounts of color; see n. 71 to *Naming and Necessity* (Kripke, 1980, p. 140).

10 Dummett contrasts the 'proper' or 'genuine' future tense with "the future tense used to express present tendencies." "The latter occurs, e.g., in an announcement of the form 'The wedding announced between *A* and *B* will not now take place.' Such an announcement cancels, but does

not falsify, the earlier announcement, and is not itself falsified if the couple later make it up and get married after all; if this were not so, the 'now' would be superfluous" (Dummett, 1972, p. 21).

11 Vann McGee (1985) presents a case where, he thinks, we are not prepared to use *modus ponens* in drawing consequences from an indicative conditional; but see Rumfitt (2013, pp. 176–178 and 185–186) for an alternative analysis of his case.

12 This result is the first 'Observation' in §3 of Humberstone and Makinson (2012).

13 In *The Logical Basis of Metaphysics*, Dummett adopts a very different approach to the problem of finding an introduction rule for negation. He works in a language with an infinite collection Q_1, Q_2, … of atomic formulae. In our notation, his introduction rule for '¬' is the infinitary rule whose underlying tuple is $< P \Rightarrow Q_1,\ P \Rightarrow Q_2,\ \ldots / \varnothing \Rightarrow \neg P >$ (Dummett, 1991, p. 295). He also proposes a cognate introduction rule for '⊥': $< / Q_1, Q_2, \ldots \Rightarrow \bot >$. In the event that the atomic formulae of the language form a consistent set, his introduction rule for '¬' allows A and $\ulcorner \neg A \urcorner$ both to be true. Similarly, his introduction rule for '⊥' allows '⊥' to be true in those circumstances. These features are surely weaknesses in his theory. Since Dummett was not a dialetheist, an account which leaves it an open matter whether there can be true contradictions must be failing to characterize logically relevant aspects of the meaning of the negation sign. Similarly, an account which leaves open the possibility that the *falsum* might be true is not capturing the intended sense of '⊥.' Dummett may be right to say that in his system "no logical laws could be framed that would entail" that not every atomic sentence can be true (1991, p. 295), but that is a limitation of his system. In a system of set/formula-or-empty sequents, the rule $< / A, \neg A \Rightarrow \varnothing >$ entails that A and $\ulcorner \neg A \urcorner$ cannot both be true, and the infinitary rule $< / Q_1, Q_2, \ldots \Rightarrow \varnothing >$ excludes the possibility that $Q_1, Q_2, \ldots$ form a consistent set.

14 It has, of course, an elimination rule: $< / \bot \Rightarrow \varnothing >$.

15 We may liberalize introduction and elimination rules to those governing set/formula-or-empty sequents while retaining the requirement that such rules must be elementary. If we do this, we shall exclude the introduction and elimination rules that Stephen Read proposes for his paradoxical zero-place connective 'bullet,' a proof-conditional Liar sentence (Read, 2000, pp. 140–142). Those who regard the bullet as meaningless will wish to retain the requirement of elementariness.

16 Rumfitt (2000). When I wrote that paper, I still thought there might be something in the Harmony Thesis, so I was concerned to show how the operational rules of my system conformed to an analogue of the harmony requirement. I no longer see any grounds for requiring such conformity.

17 See again n. 11 above on purported counter-examples to *modus ponens*.

18 Recognizing the difficulties confronting his Fundamental Assumption, Dummett briefly canvassed an alternative theory whereby every connective's meaning is given by its elimination rule. On this view, justifying the Harmony Thesis "will depend upon an inverse fundamental assumption, namely, that any consequence of a given statement can be derived by means of an argument beginning with an application of one of the elimination rules governing the principal operator of that statement, in which the statement figures as the major premiss. This assumption is open to fewer intuitive objections than the fundamental assumption on which our original justification procedure rested. It is more plausible that we derive simpler consequences from complex statements only when those consequences follow logically than that we assert such statements only when they follow logically from simpler statements we have previously accepted" (Dummett, 1991, p. 281).

Dummett's account of this alternative 'pragmatist' theory is sketchy, although Prawitz (2007), Queiroz (2008), and Litland (forthcoming) have developed it further. In particular, Litland (forthcoming, §4) corrects various mistakes in Dummett's sketch, and shows that a cleaned-up Inverse Assumption justifies precisely the intuitionistic introduction rules for the connectives, given the intuitionistic elimination rules for them. It is good to know where this approach leads. In later writings, however, Dummett came to doubt if the sort of pragmatist theory of meaning that the Inverse Fundamental Assumption requires could be coherently elaborated (see especially Dummett, 2007). In Rumfitt (2017), I identify a number of foundational problems that pragmatist theories of meaning must face, and criticize extant attempts to solve them.

19 In Steinberger's terminology, the 'I-principles' pertaining to a sentence *A* state the conditions in which a speaker of the relevant language is entitled to assert *A*. The corresponding 'E-principles' state what a speaker who asserts *A* is thereby entitled to do. (See 2011, p. 618.)

20 Peregrin (2008) argues that intuitionistic logic is the strongest logic that makes inferences explicit. He reaches this conclusion, however, by importing a number of contentious assumptions into the explanation of what it is to make an inference explicit.

21 If the introduction and elimination rules of a new connective are in harmony, will the resulting system conservatively extend the pre-existing consequence relation? Prawitz (1994, p. 374) argued not: the natural introduction and elimination rules for the truth-predicate are in harmony, but the result of adding a truth-predicate to Peano Arithmetic is not a conservative extension of it. See, however, Hodes (2004, pp. 148–150) and Steinberger (2011, pp. 635–637) for reasons to doubt whether the scope of introduction and elimination rules should be extended to encompass predicates as well as operators.

22 See especially Tennant (1987, ch. 10), which patiently untangles passages in Dummett's early writings on the topic (1975a; 1975b) that mix up the two requirements.

23 At least, he does in his book *Anti-Realism and Logic* (Tennant, 1987). In *The Taming of the True*, he holds that the introduction rules give the meanings of the connectives as they are used in *a priori* science whereas the elimination rules give their meanings as they appear in empirical discourse. Harmony is then needed to ensure that there is no equivocation between the two sorts of occurrence (Tennant 1997, p. 23). Unfortunately, I lack the space to analyze this argument here.

24 See especially p. 94: "There is another kind of equilibrium, which would be of interest even to one who refuses to acknowledge the asymmetric division of rules into those that are constitutive and those that are merely explicative of meaning. This way is to regard the rules of introduction and elimination as equally involved in specifying or constituting meaning, but to demand harmony as a *conditio sine qua non* of their doing so. The thought would be that not just any set of rules will do in order to confer determinate meaning on a logical operator."

25 It is Tennant's second thesis that justifies his claim that the correct laws of logic are confined to the rules of intuitionistic *relevant* logic. This logic yields the *least* consequence relation that satisfies his two Principles of Harmony. Litland (forthcoming, part II) shows in effect that full intuitionistic logic is the strongest logic that satisfies the two Principles.

26 This is, indeed, Hodes's position in his (2004). He holds that only introduction and elimination rules can constitute the sense of a connective (p. 147), and requires that the elimination rule should be the maximum inverter of the introduction rule and that the introduction rule should be the maximum invertee of the elimination rule (p. 156). (This amounts to Tennant's requirement of Harmony.) Hodes defines the 'basic logic' of a language to be that comprising only the sense-constituting rules for the connectives (p. 151). Given his requirement of Harmony, he takes the basic logic for English to be first-order intuitionistic logic (p. 151). However, he allows that other sorts of rule, including *EM*, are fully justified (p. 154), so that the 'total logic' for ordinary mathematical English is classical.

Hodes advances no argument for the Harmony requirement: he simply presents it as a conjecture whose implications are worth tracing out. Given that he allows the legitimacy of *EM*, though, his acceptance of the Harmony Thesis is in any case somewhat half-hearted. On his view, *EM* is a legitimate part of our inferential practice with negation. From an IRS perspective, then, it *is* part of the meaning of 'not.' I do not see the point of saying that, because *EM* is neither an introduction nor an elimination rule, it is not part of the *sense* of that word.

27 This account of what is wrong with 'tonk' is essentially that proposed by J. T. Stevenson in his reply to Prior (Stevenson, 1961). Stevenson's reply was rather eclipsed by Belnap's, which appeared the following year and started the harmony hare running. Brilliant as Belnap's paper is, I think it was Stevenson who gave the better explanation of why 'tonk' fails to have a sense.

28 In the Dewey Lectures which he delivered at Columbia University in 2002 (published as Dummett, 2006), Dummett retreated to this position. "The proponent of a truth-conditional

theory of meaning," he wrote, "must argue that [the] use [of sentences] cannot be described without appeal to the conditions for the truth of statements … To an important degree, such an argument would be correct" (Dummett, 2006, p. 29). Truth is 'indispensable' in describing how sentences are used because "a salient part of using a language is to give arguments in support of some conclusion," so that a full description of their use "needs a notion of truth, as that which is guaranteed to be transmitted from premisses to conclusion of a deductively valid argument" (2006, pp. 29, 31, 32).

References

Auxier, R. E., and L. E. Hahn, eds. 2007. *The Philosophy of Michael Dummett* (The Library of Living Philosophers Vol. XXXI). Chicago: Open Court.

Belnap, N. D. 1962. "Tonk, plonk, and plink." *Analysis*, 22(6): 130–134.

Brandom, R. B. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. B. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Dummett, M. 1959. "Truth." *Proceedings of the Aristotelian Society*, 59: 141–162.

Dummett, M. 1972. "Postscript to 'Truth.'" In *Logic and Philosophy for Linguists: A Book of Readings*, edited by J. M. E. Moravcsik, pp. 220–225. The Hague: Mouton. Page references are to the reprinting in Dummett, 1978, pp. 19–24.

Dummett, M. 1975a. "The justification of deduction." *Proceedings of the British Academy*, 59: 201–231.

Dummett, M. 1975b. "The philosophical basis of intuitionistic logic." In *Logic Colloquium '73*, edited by H. E. Rose and J. Shepherdson, pp. 5–40. Amsterdam: North Holland.

Dummett, M. 1977. *Elements of Intuitionism*. Oxford: Clarendon Press.

Dummett, M. 1978. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. 1981. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.

Dummett, M. 1990. "The source of the concept of truth." In *Meaning and Method: Essays in Honor of Hilary Putnam*, edited by G. Boolos, pp. 1–15. Cambridge: Cambridge University Press. Page references are to the reprinting in Dummett, 1993, pp. 188–201.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. London: Duckworth.

Dummett, M. 1993. *The Seas of Language*. Oxford: Clarendon Press.

Dummett, M. 2000. *Elements of Intuitionism*, 2nd edn. Oxford: Clarendon Press.

Dummett, M. 2006. *Truth and the Past*. New York: Columbia University Press.

Dummett, M. 2007. "Reply to Dag Prawitz." In Auxier and Hahn, 2007, pp. 482–489.

Gentzen, G. 1935. "Untersuchungen über das logische Schliessen I." *Mathematische Zeitschrift*, 39(2): 176–210, 405–431.

Gentzen, G. 1969. *The Collected Papers of Gerhard Gentzen*, edited by M. E. Szabo. Amsterdam: North Holland.

Hodes, H. T. 2004. "On the sense and reference of a logical constant." *The Philosophical Quarterly*, 54(214): 134–165.

Humberstone, I. L., and D. C. Makinson. 2012. "Intuitionistic logic and elementary rules." *Mind*, 120(480): 1035–1051.

Kripke, S. 1980 (1970). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, D. K. 1997. "Naming the colours." *The Australasian Journal of Philosophy*, 75(3): 325–342.

Litland, J. E. Forthcoming. "Proof-theoretic justification of logic."

Lorenzen, P. 1950. "Konstruktive Begründung der Mathematik." *Mathematische Zeitschrift*, 53(2): 162–202.

Lorenzen, P. 1955. *Einführung in die Operative Logik und Mathematik*. Berlin: Springer.

Makinson, D. C. 2014. "Intelim rules for classical connectives." In *David Makinson on Classical Methods for Non-Classical Problems*, edited by S. O. Hansson, pp. 359–382. Dordrecht, Netherlands: Springer.

McGee, V. 1985. "A counterexample to modus ponens." *The Journal of Philosophy*, 82(9): 462–471.

Mill, J. S. 1891. *A System of Logic, Ratiocinative and Inductive*, 8th edn. London: Longman.

Moriconi, E., and L. Tesconi. 2008. "On inversion principles." *History and Philosophy of Logic*, 29(2): 103–113.

Negri, S., and J. von Plato. 2001. *Structural Proof Theory*. Cambridge: Cambridge University Press.

Pagin, P. 2009. "Compositionality, understanding, and proofs." *Mind*, 118(471): 713–737.

Peacocke, C. A. B. 1987. "Understanding logical constants: a realist's account." *Proceedings of the British Academy*, 73: 153–200.

Peregrin, J. 2008. "What is *the* logic of inference?" *Studia Logica*, 88(2): 263–294.

Prawitz, D. 1965. *Natural Deduction: A Proof-Theoretical Study*. Stockholm: Almqvist and Wiksell.

Prawitz, D. 1974. "On the idea of a general proof theory." *Synthese*, 27(1–2): 63–77.

Prawitz, D. 1987. "Dummett on a theory of meaning and its impact in logic." In *Michael Dummett: Contributions to Philosophy*, edited by B. M. Taylor, pp. 117–165. Dordrecht, Netherlands: Martinus Nijhoff.

Prawitz, D. 1994. "Review of *The Logical Basis of Metaphysics*." *Mind*, 103(411): 373–376.

Prawitz, D. 2007. "Pragmatist and verificationist theories of meaning." In Auxier and Hahn, 2007, pp. 455–481.

Prior, A. N. 1960. "The runabout inference-ticket." *Analysis*, 21(2): 38–39.

de Queiroz, R. J. G. B. 2008. "On reduction rules, meaning-as-use, and proof-theoretic semantics." *Studia Logica*, 90(2): 211–247.

Read, S. L. 2000. "Harmony and autonomy in classical logic." *Journal of Philosophical Logic*, 29(2): 123–154.

Read, S. L. 2010. "General-elimination harmony and the meaning of the logical constants." *Journal of Philosophical Logic*, 39(5): 557–576.

Rumfitt, I. 2000. "'Yes' and 'no.'" *Mind*, 109(436): 781–823.

Rumfitt, I. 2013. "Old Adams buried." *Analytic Philosophy*, 54(2): 157–188.

Rumfitt, I. 2017. "Tempered pragmatism." In *Pragmatism in the Long Twentieth Century: Proceedings of the 2014 Dawes Hicks Symposium*, edited by C. Misak and H. Price. London: British Academy.

Schroeder-Heister, P. 1984. "A natural extension of natural deduction." *The Journal of Symbolic Logic*, 49(4): 1284–1300.

Steinberger, F. 2011. "What harmony could and could not be." *Australasian Journal of Philosophy*, 89(4): 617–639.

Stevenson, J. T. 1961. "Roundabout the runabout inference-ticket." *Analysis*, 21(6): 124–128.

Tennant, N. W. 1987. *Anti-Realism and Logic: Truth as Eternal*. Oxford: Clarendon Press.

Tennant, N. W. 1997. *The Taming of the True*. Oxford: Clarendon Press.

11

Meaning and Privacy

EDWARD CRAIG

1 Introduction: The Two Questions and their Consequences

It has been widely held that certain states of sentient creatures are private, in the technical sense that their nature cannot be known by anyone other than the subject who experiences them. For instance, the phenomenal quality of perceptual states has often been seen in this light. "I know you call that color by the same name, but I can't know whether you see it in the same way as I do" is a position familiar to most students of philosophy, both amateur and professional. Involved, clearly, are issues in both the philosophy of mind and epistemology, but it is not the purpose of this chapter to go into these in depth; for the moment the reader should assume – or at least be prepared to entertain the hypothesis – that there are indeed states which are private in the sense defined. We may call them epistemically private items (EPI).

A general question arises about the role, if any, of such EPI in the meaning of language. A highly influential tradition makes the meaning of a word depend on the nature of the "idea" associated with it, whilst treating ideas as items before the consciousness of speakers and their hearers, hence as strong candidates for epistemic privacy. Much recent argument, on the other hand, denies that the epistemically private can have any such part to play, and that it is precisely the fact of its privacy that rules it out of semantics.

The debate thus broadly characterized focuses not on one question, however, but two, close enough to be conflated by the unwary, yet quite different enough for all hope of clarity to be gone if they are not carefully distinguished. One concerns the semantics of the "public" language, the one in which we communicate, or apparently succeed in communicating, with each other; the other is the notorious question about the possibility of a "private" language, that is to say a language used by a person for the sole purpose of communicating with their own (later) self, and in principle unusable for communication with anybody else. We shall have to ask:

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

(1) whether the nature of our EPI affects, or can affect, the semantics of a natural language used for interpersonal communication
(2) whether there can be a private language, in which a person records facts about their EPI for their own later information.

The content of the first of these questions is easy enough to grasp, even if the arguments that are brought to bear on it may be less so. But a "private" language sounds like a highly artificial construction, and a little time must therefore be spent inquiring just what such a language is supposed to be.

The *locus classicus* is Ludwig Wittgenstein's *Philosophical Investigations*, §243:

> The words of this language are to refer to what can only be known to the person speaking: to his immediate private sensations. So another person cannot understand the language.

As an introduction to the concept of a private language this passage is not without its difficulties. The first sentence appears to commit itself to at least a limited skepticism about the contents of other minds, though it is probably better read as a definition of the concept of privacy, with the plausible assumption added that if there is anything private in this sense, the phenomenal quality of states of consciousness will be so. The second sentence implicitly brings in another assumption: that A can understand B only if he *knows* the nature of the objects B is referring to. We shall later see that this is open to question, with the result that it may make a difference whether we take "private language" to be defined by the first sentence of the quotation or the second. By no means all the literature on the topic pays due attention to this possibility.

Since the assimilation of Wittgenstein's *Philosophical Investigations* (1953) it has been common, indeed almost standard, to answer both questions in the negative: EPI have no role to play in the semantics of any public language; and there can be no such thing as a private language. Moritz Schlick, in a series of lectures delivered in London in 1932, in effect had made both these claims. More recently Michael Dummett (Dummett, 1973) has offered reasons in favor of the negative thesis about EPI and public language; the extent to which these arguments are new, and the extent to which they reformulate or adapt points from earlier literature, will be considered in what follows.

The two questions have not, in the main, been pursued for their own sakes, but rather because negative answers to them have been thought to entail important consequences in the philosophy of mind, the theory of knowledge, and metaphysics.

One major question affected, bridging the first two of these areas, is that of skepticism about knowledge of other minds – which dissolves as unstatable if the answer to (2) is negative. Another, obviously central to the philosophy of mind, is that of the true meanings of *prima facie* EPI-words, which is obviously much affected by a "No" to (1).

(It might be thought that (1) was capable of resolving or dissolving skepticism about other minds without the assistance of (2). For if those ("private") aspects of our mental states on which the skeptic may with good prospects focus attention have no effect on what any word of our common language means, it would seem to follow that we cannot say anything to each other about those aspects, hence cannot pose the question of whether they are the same in others as they are in ourselves. But this overlooks the fact that, without a negative answer to (2), it still remains a possibility that each of us can formulate for ourselves the question "Are their EPI like mine?" and then doubt whether we can know the answer.)

In epistemology, a "No" to (2) poses problems for those "foundationalist" theories which purport to ground knowledge on beliefs about items that might be taken to be EPI; also for doctrines which "analyze" propositions about the *prima facie* public into propositions about the *prima facie* private, such as phenomenalism. It could be said that so long as (2) only speaks of what can/cannot be *said*, it doesn't affect these doctrines (which only require assumptions about what can be *thought*); but – apart from the fact that the implied distinction between being sayable and being thinkable may prove troublesome – one must consider that we shall find arguments being deployed against a private language which attack the notion of thought about one's EPI in the first instance, and language only derivatively.

Further (and more recently), the answer to (1) has been suggested (by Dummett) to be of critical importance in a central question of ontology: that of the debate between realists and anti-realists. (See Chapter 20, REALISM AND ITS OPPOSITIONS.) A negative answer, in Dummett's view, leaves the anti-realist in the ascendant. Later on we shall briefly consider whether (1) really is central for the realism versus anti-realism debate.

2 Private States and Public Language: The Possibility

Michael Dummett has denied that EPI can play any role in the semantics of the public language. For this view he advances a group of three closely related arguments, which we may call respectively the arguments from Communicability, from Acquisition, and from Manifestation.

Suppose that what I mean by some expression of the public language is affected by the nature of certain states epistemically private to me. Then, argues Dummett, that expression has a meaning which I will not be able to communicate to anyone else. Because the states are private, nobody else can know their nature, hence neither can they know what I mean by that expression. Others do not understand it, therefore it is not, after all, an instrument of the public language, which is by definition a vehicle of interpersonal understanding.

The Argument from Acquisition asks us to consider the position of learners acquiring the use of a language, some of whose expressions depend for their meaning on the nature of their teachers' EPI. Since *ex hypothesi* this is something they cannot know, they cannot make the associations needed to know what the expressions mean; in other words, they cannot acquire a grasp of the language.

The Argument from Manifestation looks at the same situation from the point of view of those who, like the teacher, have to judge whether or not the learners now use the expressions of the language with the accepted meanings. But if these meanings are affected by speakers' EPI, there is nothing the learners can do to "manifest" their understanding, no way in which they can let the competent speaker know that they have learnt their lesson properly.[1]

In short, if EPI affect meaning, then what anyone means is unknowable to anyone else, and language cannot serve mutual understanding or communication. Hence it is not a public language. Anyone who thinks that interpersonal communication is of the essence of language will be prepared to drop the word "public" from that sentence, and say that such a "language" is not a language, *tout court*; but so long as the possibility of a private language is undefeated we should stop at the narrower conclusion.[2]

It will be noted that all these arguments turn on the same point: if the meaning of other speakers' expressions depends on the nature of their EPI it cannot be known. They are thus very close relations of an argument that has frequently been directed at the classical empiricist

theory of meaning. That theory equated meanings (and thoughts) with image-like items entertained before the mind's eye, that is to say with some of the strongest candidates for epistemic privacy; so it had the intolerable consequence that nobody could know what anyone else meant or thought.

The most direct approach to the arguments from Communicability, Acquisition, and Manifestation will therefore be to ask two questions. First, is it true that if EPI have a role in semantics there can be no interpersonal knowledge of the meanings of the expressions in whose semantics they figure? Second, if that be so, how much does it matter? – is it the catastrophe that the arguments imply it to be, or can a supporter of EPI-semantics just take it in his stride?

The answer to the first question must be yes, since it follows straight from the definition of an EPI, given the principle that if A is an essential constituent of B, the nature of B can be known only if that of A is knowable too. But there is a complication which should be considered, if only to avoid being confused by it. This is the thought that until we have inspected the cases for and against skepticism about Other Minds, the traditional "Argument from Analogy," and the like, it must remain an open question whether we can know what someone else means, even if their meaning is determined, or partially determined, by their inner states.

In one respect the introduction of this thought at this point is just an irrelevance, for it is aimed not so much at the thesis that EPI can play no role in public-language semantics as at the question of whether there are any EPI at all, any facts about persons which are epistemically private in the defined sense. But in another respect, the matter of the *application* of the thesis rather than its internal logic, it is very much to the point, since it asks whether there is in fact anything at all about us which the thesis would, if true, exclude from semantics and pronounce incommunicable. We shall return to this later, though only briefly. Responses to skepticism about other minds are very varied, as are the conceptions of the mental which they trade upon or promote, so that a serious discussion of these possibilities would lead us too far away from our topic of the relation between private states and meaning.

So let us for the moment accept the obvious affirmative answer to our first question, and move to the second. One reason for announcing a catastrophe sufficient to discredit the position it follows from runs: Understanding someone is knowing what they mean. So if I don't know what you mean then I don't understand you. Hence any theory which makes the meaning of expressions of the public language dependent on EPI in effect denies that we understand each other when using these expressions; and that means that, by definition, they are not elements of the public language at all.

It may be doubted, however, whether understanding someone is necessarily to be equated with knowing what they mean. Accounts of the concept of knowledge vary, and some allow that knowledge exists where others deny it, so that any casual equation of understanding with knowledge of meaning needs more rigorous investigation. We may approach this, whilst keeping in touch with our main question, by considering the possible effects of what I shall call "Burke's Assumption":[3] we naturally assume that others have, in broadly similar external circumstances, broadly similar internal states. In so far as the latter are epistemically private there is no question of our knowing whether, or when, the assumption is true, but nevertheless it is one we all naturally make. Under exceptional circumstances (Fred can't see which traffic light is on, Mabel can't tell whether that is a trombone or a flute – whatever the incentive we offer them to get it right), we start to adjust our beliefs, but this is a departure from our first, firm inclinations. And let us assume further, that these assumptions are in fact mostly right.

Given Burke's Assumption, it will often be the case that speaker and hearer, teacher and pupil have the same, true beliefs about the qualities of each others' EPI; and hence there will no barrier in principle, and usually none in fact, to each arriving at the same, true beliefs about what the other means, even though their EPI affect their meaning. Is this enough for the operation of a public language in the semantics of which EPI have a role? And if it is enough, what of the assumptions that were necessary to allow us to reach this position?

To begin with the first question, we still have to admit that none of our speakers *knows* that the others mean what he takes them to mean. But this would not seem to prevent anyone from understanding anyone else. There is no particular reason to think that to understand correctly one must know that one understands correctly. And there seems to be no reason to deny that such true beliefs about the meanings that others attach to their expressions are enough to constitute understanding. After all, speakers will be expressing the thoughts that their hearers confidently take them to be expressing, and why should that be held to fall short of understanding them? They do not have any guarantee of mutual understanding, but why should there not be understanding without a guarantee?

Perhaps the objection is a rather different one, however. Whatever may or may not be required for understanding, it remains the case that we *do* know that we understand each other. And that rules out all theories which cast publicly unknowable items in any essential semantic role.

With this we are back at the point we touched on earlier, and the proponent of EPI-driven semantics has two lines of reply. One would be to adopt, at least as a defensive measure, some "externalist" account of knowledge such as reliabilism, and then point out that if we do make such assumptions as Burke suggested, and these assumptions are, in the main, correct, then our methods of coming to beliefs about other people's EPI are reliable, and the beliefs accordingly are knowledge; the supposed EPI aren't private after all. It still applies, admittedly, that if there are any states of persons about which others have no reliable ways of coming to true beliefs, then they really are EPI and can have no effect on the public language; but whether there are any, and if so which, becomes a very obscure and perhaps from the very nature of the case uninvestigatable question.

The trouble with this blocking response is that it relies on a particular, by no means uncontroversial, account of the concept of knowledge. And even if this account be accepted, the objector to EPI-semantics might well not agree that his worries had been adequately addressed. What the objector is really saying that there must surely be (quite independently of the correct analysis of "S knows that p," even if such a thing exists) is some good reason – something which we can see to be a good reason without needing to know already that the beliefs are true. And it is precisely this which there cannot be, with respect to our belief in mutual understanding, if EPI have any effect on meaning.

This argument will be as strong as the claim made by our objector, that there must be some reason (of an "internalist" kind) for this belief. How strong is that? Perhaps not very strong, for there is a plausible line of thought that sheds a good deal of doubt on it. Reaching a belief by reasoning to it from premises antecedently believed is a slow and uncertain process, easily thrown off course. Seeing that for practical purposes many beliefs are needed quickly, and with the degree of conviction necessary to give rise to immediate and decisive action, evolutionary development is not likely greatly to have favored it over "blind," as it were mechanical methods of acquiring beliefs. Once a certain level of sophistication has been reached, the power of reasoning, used at the right time and place, may become a most valuable way of extending our stock of beliefs; but there are no grounds for thinking that it

took much part in the beginnings of human mental activity when our basic types of belief and methods of forming them were being developed. What happened then was that we grew some successful psychological hardware. But that means that when we now consider some class of beliefs (such as those involved in Burke's Assumption), we are not entitled to *assume* that they must be retrospectively certifiable by any process of rational inference; they may turn out to be so, and then again they may not. And one thing we should certainly not do is think that they must be rationally certifiable just because we find them so convincing.

If allowed, this argument seriously weakens the claim that, since it would mean that in many cases we could give no good reason for thinking that we understand each other, EPI can have no part in semantics. For perhaps we can give no such reason. But some may think that the argument ought not to be allowed, after all. They can point to the unquestionable fact that there are many beliefs that we almost certainly would not have unless we could give reasons for them, and then they *may* be able to argue that the belief in mutual understanding is likely to be one of them. How to turn this possibility into a concrete proposal, however, is not obvious; and the present writer knows of no published attempt to do so.

Those who think that the nature of speakers' EPI may, without disaster, be assigned a role in the meaning of expressions of the public language thus seem to have the better of this phase of the argument, if only perhaps for the time being: at least until further arguments are brought, they can stave off the objection that their proposal puts our confidence in mutual understanding beyond the scope of rational support. But there is another, quite different line of attack against which they will also have to defend themselves.

The idea was to achieve mutual understanding, without banishing EPI from semantics, by relying on Burke's Assumption. More explicitly, we were to suppose that we all have a strong tendency to attribute to others EPI like our own under like circumstances, and that this tendency mostly leads us to true beliefs. The first type of objection focused on the *consequences* of this supposition, and turned on the question of whether they were unacceptably skeptical or not. The second type of objection is directed at Burke's Assumption itself: does it really make sense? Whether true or false, Burke's Assumption makes essential use of the notion of interpersonal comparison of inner states: we generally assume, it says, that those of others are similar to our own. It has been argued, however, and in more than one way, that this notion is in fact incoherent.

One such line of argument comes, unsurprisingly, from verificationism. Any doctrine which links meaningfulness at all closely to verifiability, whatever may be the exact nature of the link, is bound to find difficulty in alleged comparisons between the EPI of different persons. For since no subject can, in principle, have knowledge of both terms, any such comparison is as good an example of unverifiability-in-principle as can be found in a sentence that neither introduces "nonsense" vocabulary nor flouts any basic rule of grammar. If verificationism doesn't exclude this as meaningless, how could it exclude anything that we wouldn't all exclude anyway?

Nothing can safely be concluded from this, however. Notoriously, it proved extremely difficult even to formulate the verification principle in any way fully satisfactory to its proponents; so it is hardly surprising that the few published attempts to argue for its acceptance have turned out inconclusive. In any case, those likely to accept directly verificationist lines of thought are nowadays far fewer than was the case a generation ago. (See Chapter 4, MEANING, USE, VERIFICATION.) A much more fashionable assault on Burke's Assumption issues from Wittgenstein's discussion of rules and what it is to follow them.

Saul Kripke (1982) has offered an argument inspired by Wittgenstein's writings on rule-following, if not actually to be found in them. Suppose that a speaker uses a word on some occasion in the same way, that is to say with the same meaning, as he has used it on an earlier occasion. There must be something in virtue of which his present use of the word is consistent with previous uses (rather than having another, new meaning on this occasion). What could this something be? It cannot be either a publicly observable fact about him or his behavior, or a fact about his inner mental state.[4] So it cannot be a fact solely about *him* at all, but must include something about the behavior of other speakers of his community, to the effect that they would speak as he does in these circumstances, or regard his way of speaking as correct. The notion of sameness or consistency of meaning therefore demands the existence of some communal practice and is illusory without it.

If this be true (a very difficult buck which I here thankfully pass), then there is certainly a *prima facie* threat to the participation of EPI in semantics. I think we may take it that meaning is not a property of totally isolated utterances, but arises because expressions, and ways of combining them, are used consistently in accordance with specific rules. It follows that something can be a factor in the meaning of an expression only if it relates in a consistent way to the use of that expression, which means that, whatever it is, there must be a coherent notion of its being "the same thing again" or "another of the same sort." Is there any such notion where EPI are concerned? On the assumptions we are making about the outcome of the rule-following debate, that reduces to the question of whether, as regards the description of EPI, there is such a thing as communal agreed practice. And the temptation is to say no, precisely because they are private.

Before succumbing to it, however, there is a somewhat convoluted line of thought which we need to follow through. To start with, note that the principle about communal practice will surely have to be hypothetical in form, a matter of how others *would* describe something if they were well placed to do so. Otherwise we invite the result that if some potholer is the only person ever to see a certain underground rock formation, he cannot possibly describe it, either correctly or incorrectly. So perhaps our question should be not: How do others describe my EPI? but rather: How would others describe my EPI if they were well placed to do so?

Now this might seem the right moment to say that others never are or could be well placed to describe my EPI, precisely because they are private to me. So our conditional ("If others were well placed …") isn't assessable, even in principle, and the original temptation beckons again. Easiest would be just to give in to it; but we shouldn't, because if Burke's Assumption is true, then we frequently are well placed to describe other people's EPI. Perhaps we aren't *as well placed as they are*, but why should that be necessary? For after all, the requirement we allowed ourselves to start from, provisionally accepting it as a consequence of the rule-following debate, was the need for "communal practice." It would be a further thing to demand that this communal practice be based on *knowledge* of the items being described. Whether it could be justified or not could only be settled by a detailed scrutiny of the arguments about rules; but since it has not been established uncontroversially that even our provisional assumption really does follow from them, the prospects for a yet stronger version specifying knowledge as an essential basis of the communal practice must be quite doubtful.

It appears likely, then, that this type of attack on Burke's Assumption merely begs the question; its pivotal claim, that no relevant communal practice exists, can be made only when it has *already* been shown that Burke's Assumption is false, or incoherent. The possibility that EPI may have a part to play in the semantics of a public language remains open.

Another line of attack begins with the arguments against the possibility of a private language. If these arguments show that we cannot communicate with ourselves (or perhaps it should be "our later selves") about our EPI, then surely *a fortiori* we cannot communicate with others? But even this question is cloudy. It was pointed out earlier that the *locus classicus*, Wittgenstein's *Philosophical Investigations*, §243, actually offers two definitions of a private language. Wittgenstein appears to have assumed that a language in which someone speaks of "what can only be known to the person speaking; … his immediate private sensations" would necessarily be one which others could not understand. But our argument so far suggests that this may well be mistaken, in which event the two definitions fall apart and clarity demands that we look at them and their consequences separately.

Fortunately we can quickly clear the air, at least to some extent. If a private language be defined as "a language which only one speaker can, in principle, understand," then its alleged impossibility can have no effect at all on the question of whether EPI can figure in the semantics of a language which many people can understand, *unless* we take it that such a language would refer to the speaker's EPI, and that that is the ultimate reason for its impossibility. Otherwise the notion of an epistemically private item will simply not get a foothold in the logic of the argument, and our investigation reaches a dead end. In effect, we find ourselves forced back to the definition of a private language in terms of EPI.

So: if a private language, understood as one in which speakers refer to their own EPI, had been shown to be impossible, wouldn't it follow that EPI had no role in the public language either? But still the mists won't disperse, because we have to answer that we can't yet say: it will depend on just why a private language is impossible. If the agreed reason is that for EPI there is no legitimate notion of "being the same" or "being of the same kind," then will this not affect the private and public questions equally?[5] Perhaps, but only if the reasons for declaring there to be no such notion apply to the public case as strongly as to the private case. If they are verificationist reasons, then surely they do; it is hard to imagine how interpersonal verification could be thought possible when intrapersonal verification was not. But what if they are reasons drawn from the rule-following debate, to the effect that meaningfulness calls for a social practice, in other words a multiplicity of speakers? Wouldn't they have force only against the private language, leaving the public language, where by definition there is more than one speaker, untouched?

If we find that last line of thought convincing, however, then that can only be because we are still conflating the two definitions of a private language. The definition we are now supposed to be concentrating on doesn't say anything about a private language having only one speaker; it defines a private language as one in which you refer to your own EPI. And the crucial point is that until our earlier arguments about understanding and Burke's Assumption have been decisively refuted, it remains possible that the public language may be a private language as well; for if the public language permits us to talk to each other about our EPI, then why not also to ourselves? The idea may sound paradoxical, but only if we are still caught up in the second half of Wittgenstein's unfortunate double definition, and so feel that a private language *must* be a language with only one speaker.

Contrary to much recent thought, then, it begins to seem quite possible that there are no conclusive reasons for banning EPI from the factors that can give expressions of the public language their meaning. But that only brings us to the next two questions: Provisionally accepting that it is *permissible* to do so, should we ever *actually* involve EPI in our account of meaning? And what difference will it make if we do?

3 Private States and Public Language: The Effects

The immediately obvious candidates for EPI-affected semantics are expressions which purport to describe sensations (like "itch" or "headache"), and those standing for properties at least plausibly thought of as powers to produce sensations of certain kinds (such as "red" or "shrill"). Here it is tempting to think that what a speaker means depends on the phenomenal quality of his experiences; and most people would take the view that, whereas there may be other examples, if there are to be any at all then these must be amongst them.

But even here there can seem to be two options. One is to give in happily to the temptation. The other is to stick to the idea that the function of the public language is the adjustment and coordination of behavior, and that anything surplus to that, whatever its standing, is no part of linguistic meaning; so that provided we agree about what is to be called "red," stop at red lights, anticipate sweetness in red apples, we agree on the meaning of "red," and would do even if our respective visual experiences were quite different. Plenty that is epistemically private is going on, on this view, when we see red lights or have headaches; but whatever its significance for human life, it doesn't affect the meanings of any of the expressions we use to talk to each other.

Coming at the present stage of the argument, however, this seems unmotivated. If we have already agreed that nothing bars EPI from playing a semantic role in principle, isn't it merely doctrinaire to insist that they never do so in fact? The restrictive view of the function of language described in the last paragraph has usually stemmed from a decision that only what is public can be of any import in semantics, and there is no obvious reason to stick to it if that decision is itself in doubt. Admittedly it is sometimes useful to distinguish between understanding a speaker's words and understanding the speaker, in the sense of knowing what it is like for them to be in the situation that their words describe: you may know exactly what is meant by my utterance "There's a snake coming toward me," whilst having no idea how I feel about it. But what principally makes the distinction useful in this case is that what I have said may be true *however* I feel about it; so it has no application to a case in which what I am doing is describing my feelings.

Anyone prepared to go this far, and to allow that private states may affect, indeed be objects of, public discourse, is already quite a long way away from what has become, since the *Philosophical Investigations*, more or less the standard position. But most recent interest in the question about the semantic role of private states arises from the belief, advocated by Michael Dummett, that a very deep and general metaphysical issue depends on it: whether it is permissible to take a realist view of the world.

That issue is naturally, if vaguely, understood in terms of the world's dependence on, or independence of, human styles of thought and methods of investigation. But Dummett would have it understood, at least in the first instance, as a question about the right form for a theory of meaning: should the meaning of a sentence ultimately be characterized in terms of its truth-conditions, or in terms of the conditions under which we regard it as assertible? We need not here concern ourselves with Dummett's reasons for recommending this question as a fruitful entrance to the debate about realism, but can concentrate on his view that a decision against any possible role for EPI promotes the "assertibility conditions" approach to semantics. (See Chapter 20, REALISM AND ITS OPPOSITIONS.)

Grasp of meaning, once any part is denied to EPI, must consist in the capacity for some kind of publicly accessible behavior. So argues Dummett, and he goes on to say what that behavior must be: recognition of the circumstances under which the sentence can properly

be asserted. We may think we understand sentences whose truth-conditions, if they obtain, we cannot recognize as obtaining. If so, we delude ourselves, since under those circumstances our understanding could not be manifested, but would have to consist in some epistemically private feature of our minds, in breach of the principle that EPI have no legitimate business in semantics. Assertibility conditions, not truth-conditions, must therefore be primary in a theory of meaning.

This argument certainly has some force, once we accept the ban on the private from which it starts. But two corners have to be negotiated before it can be fully convincing, and they should at least be signposted here. First, we have to consider whether the explicit recognition of the fact that certain conditions obtain really is the *only* sufficient way of manifesting grasp of the meaning of a sentence. Second, it may be asked whether such "explicit" recognition really is as publicly accessible a phenomenon as the proposed use of it demands.

Neither question is easily resolved. The first is obscured by the point that manifestation must mean manifestation to someone, in this case other speakers of the relevant part of the language, which is to say other human beings familiar with the subject-matter. But do we not then need to know in advance that we cannot entertain unverifiable thoughts? For if we can, may there not be numerous ways in which someone can manifest to others that he is thinking some such thought, and using a certain sentence to express it? If this is a thought we can entertain, perhaps we also have a shared pattern of reactions to it, a pattern that can signal to others that a speaker is indeed expressing it.

The second question raises the problem of the nature of intensional states. Recognizing something, in this case the obtaining of certain truth-conditions, is not just a matter of assenting *when* they obtain, but of doing so *because* they obtain, and because *they* obtain rather than because of some other conditions which accompany them. Thus the notion of recognition brings with it something which might be called the "perspective" of the subject who does the recognizing; and it is not obvious that this can be accounted for in terms restricted solely to publicly available features of the subject and the situation. Perhaps it can, but complex issues are involved.

So it should not be thought that anti-realism in the theory of meaning follows directly from the thesis that EPI can play no part in semantics. On the other hand, it would be just as bad a mistake to think that if that thesis is shown to be groundless we can at once help ourselves to meaning-theoretic realism. True, it would then have been shown that if we can entertain thoughts about states of affairs whose existence we could not in principle detect, those states of affairs could be the truth-conditions of sentences of a mutually understandable public language, regardless of whether entertaining such a thought called for the occurrence of certain private states. But we would still not have addressed the question of what it would be to think such a thought; nor would that question necessarily be any easier just because we were allowed to appeal to EPI in answering it.

4 The Possibility of a Private Language

We now return to the second question broached at the beginning of this chapter: the possibility of a language in which a subject can, comprehensibly at any rate to himself, express thoughts about his own private states. This question is for historical reasons now inseparably connected with certain passages from Wittgenstein's *Philosophical Investigations*, and any treatment of it must take account of them; but they will be used

here simply as the obvious door to the debate, and no attempt will be made to decide any of the trickier questions of Wittgensteinian exegesis.

It was remarked in §1 of this chapter that Wittgenstein introduces the notion of a private language with a double definition; its words are to "refer to what can only be known to the speaker" (his "immediate private sensations") – and it is a language which no one else can understand. It will be obvious from §2 that I would wish to drop the second clause from consideration as causing far more trouble than it is worth. We shall concentrate on the idea of a language with terms that refer to a speaker's own EPI, and the crucial question is whether even the speaker himself could understand it, whether it really could be a language even for him.

But that way of putting it, though it has become standard, still does not reliably capture quite what most philosophers have had in mind when thinking about private language. What is essential is not so much the idea of terms that have the speaker's EPI for their *reference*, but rather of terms whose sense or meaning depends in some degree on the nature of the EPI of the person using them. It is, of course, true that some such terms, by virtue of their meaning, may well be usable to refer to the speaker's EPI; and it is also true that Wittgenstein's most famous example is about the (purported) use of just such a term to record recurrences of a particular type of sensation (see Wittgenstein, 1953, §258). Nevertheless, the wider formulation just suggested will serve us better. It covers more possibilities; it gives a closer parallel with the question about EPI and public-language semantics discussed in earlier sections; and the arguments of the private-language debate as actually conducted apply to it at least as well as they do to the version which, by following the standard translation of *Philosophical Investigations*, §243, makes the notion of reference sound primary.

All who argue the impossibility of a private language, and all those who defend private language against them, seem agreed on one central point: that meaningfulness requires the rule-governed, or at the very least consistent, use of a symbol; so that whatever it is about the use of a word which determines its meaning must be capable of *recurrence*, of being the *same again*. That principle has an obvious consequence for the words of a private language: epistemically private items, or those aspects of them which affect the meanings of the private vocabulary, must be the sort of thing which can repeat: it must be possible for there to be *another thing of that kind*. And this simple commitment is, according to various arguments, the Achilles' heel of the idea of a private language.

One such line of argument makes appeal to verificationism. Whether it is to be found in Wittgenstein is a matter of controversy, but there is little room for doubt that verificationism was at least one of the planks on which Schlick rested certain negative claims about private language (though he did not use that expression) in lectures given in London in 1932 (see Schlick, 1938, pp. 177–179). The critical question for him was, anticipatably, whether a color seen today was of *the same* shade as one seen yesterday. What did it mean? That was to be determined, he said, by looking at the way in which an answer could be tested. So long as we were allowed to resort to such things as the opinions of other people, the persistence of the colored object, and the empirically determinable probability of its having changed color, no special problem arose. But it was quite otherwise, Schlick held, in a case in which there was nothing to appeal to beyond the memory of the person making the judgment. In that case we should

> have to declare it impossible to distinguish between a trustworthy and a deceptive memory; we therefore could not even raise the question whether it was deceptive or not; there would be no

> sense in speaking of an "error" of our memory…. I recall it so, and that is final; in our supposed case I cannot go on asking: do I remember correctly? for I could not possibly explain what I meant by such a question. (Schlick, 1938, p. 179)

Not even in communication with oneself, Schlick concluded, can words convey the nature of a private experience.[6]

Now in so far as this argument rests on the principle that the meaning of a statement is its method of verification, few philosophers nowadays will rush to endorse it; even fewer in so far as it is felt to rest on an application of the "Picture Theory of Meaning."[7] And, as all readers of Wittgenstein will have recognized, the above passage bears a striking resemblance to parts of the *Philosophical Investigations*, striking enough to raise the question whether Wittgenstein really added anything new. But it is widely held that he did: that the *Investigations* contain an argument to the impossibility of a private language that makes no use of either the picture theory or of the verification principle.

This argument is usually located in §258, where Wittgenstein asks how the user of the private language is to give meaning to its signs. An inward ostensive definition, with the attention concentrated on the relevant private item, will be just an idle ceremony of no semantic consequence unless it brings about consistent usage: the "speaker" really does thereafter apply the sign correctly, that is to say in connection with EPI of *that* type. The trouble is that

> in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about "right."

In that event, Wittgenstein leaves us to conclude, there is no difference between a language and what merely seems to its "speaker" to be a language. And that only means that here we can't talk about a language. What vitiates private language is the collapse, in the case of EPI, of the distinction between "seems" and "is."

A distraction at this stage is the notion, introduced in §258, of an ostensive definition. Earlier in the *Philosophical Investigations* (see Wittgenstein, 1953, §28 and following) Wittgenstein discusses the business of giving meaning by ostension, arguing that its effectiveness depends on a great deal of cooperation from the recipient of the definition, who needs the right antecedent mental "set" if he is to discern what kind of thing is being pointed out as an example of the term. It is often suggested, not implausibly in view of §§257–258, that these thoughts about ostensive definition are part of Wittgenstein's weaponry against private language.

If they are, it is because, in Wittgenstein's opinion, ostensive definition can work only under conditions which the private linguist doesn't satisfy. Such at least seems to be the message of §257: "a great deal of stage-setting in the language is presupposed." And the implication of the last sentence of the paragraph is that the stage-setting in this case has to come from the public language, in particular from its use of the word "pain."

This raises a little swarm of questions. In the first place, is what is at issue here the possibility of a private language as, following *Philosophical Investigations*, §243, we defined it earlier? Or is it rather the possibility of having a private language whilst not speaking any public language? For it looks as if the most that §257 will show is the impossibility of the latter, and not of the former. (If the would-be private linguist does have the use of the word "pain" in public English, why should he not build on it a term designed to express the particular character of certain pains of his?)

Second, we should note that though ostensive definition calls for prior "stage-setting," it cannot always call for prior stage-setting *in a language*, since otherwise we could never get started. So the claim that our private linguist requires part of the public language as his stage, rather than just a particular mental set, needs special argument and shouldn't be accepted without it.

Third, we should notice the way in which §257 begins: we are to imagine a special situation, in which human beings have pains but show no outward signs of having them. Now it is quite reasonable to suppose that, under those circumstances, that part of the public vocabulary could not be taught, and so would not exist. So if there were good grounds to think that "stage-setting" of that kind would be necessary for the inner ostensive definition to work, then a private language might well turn out to be impossible. But that would have been shown to hold, we have to remember, only for the imaginary circumstances posited at the beginning of the paragraph. We could then react in two ways: we could either say that this shows nothing at all about the impossibility of a private language under the actual conditions of human life, or we could modify our understanding of what a private language is supposed to be.

The necessary modification would be to think of an EPI not just as something the nature of which could be known only to its experiencing subject, but to require also that there should be no outward sign of its occurrence. Such an extra load on the concept of an EPI would render the thesis that private language is impossible weaker and very much less important. This can be seen by reflecting that it could be accepted by the most unreconstructed Cartesians, provided that they were prepared to say that some outwardly observable feature of the material body went along with every inward feature of the mental substance of the mind, and that they would not then have to alter their views about the nature of the mental in any way at all. It would therefore be much more to the point if we could show the impossibility of private language on the old, unrevised concept of the epistemically private.[8]

Back then to the argument from the collapse of the seems/is distinction. Why is it held to have collapsed? Why shouldn't whether this EPI of mine really is of the same kind as that of yesterday be one question, whether it seems to me to be the same another? One line, we have seen, is to say that the only way to verify an answer to the first question is to ask the second – and then conclude that the two questions can't, after all, be distinguished. No more about the verificationism of that argument; but it is worth asking whether its other premise ('the only way to verify an answer to the first question is to ask the second') is true.

It could be said that the only way to verify an answer to any question at all is to find out whether it seems to us to be the right answer. But for most questions that will be true only if "seeming to us to be the right answer" is allowed to describe the result of a complex procedure in which several avenues of inquiry lead us to the same point. (What we seem to see coincides with what we seem to hear, with what Fred and Mabel seem to be telling us, and with our memory of what is normal under the circumstances seeming to obtain, for instance.) It is this that gives the seems/is distinction content: it becomes possible that something might seem, by one investigative route, what it turned out (on the witness of the other routes) not to be. And what causes the trouble for the private linguist is that in the case of an EPI there is none of this complexity, only a once-and-for-all judgment, unanswerable to any further investigation, that it seems to be of the right kind. But is that true? Must all judgments about EPI have this "one-track," structureless character?

Some writers have thought not, but suggest that there is no reason of principle why EPI should not fall into patterns and exhibit regularities (see, e.g., Harrison, 1974, ch. 6,

especially §37). One might add that there is a powerful *de facto* reason why they should: since they are our perceptual states, they must exhibit all manner of regularities if we are to perceive a stable and regular world, as of course we do. And if so, there will be more than one question which the private linguist can ask when trying to decide whether a particular item was of a certain type or not. Besides just "Did it strike me as being of that type?" there is also "Did it seem to be accompanied by the items that usually – as it seems to me – accompany items of that type?" A response to this would have to take one of two courses: either to retreat, saying that the argument only applied to such EPI as don't fall into any such patterns, whilst allowing a private language to encompass all that do; or to argue that the existence of such patterns would have no tendency to reinstate the seems/is distinction.

If we are hoping for a robust version of the anti-private-language thesis, only the second option will be of much interest. Its proponents must tread carefully, lest their reasons for denying the seems/is distinction in the private case get out of hand and threaten the distinction for public objects, thus undermining public language as well. The threat is serious enough, since it might well be thought that the only thing that enables us to make a distinction at all between seeming to be and actually being is the existence of *various ways* in which a given proposition may seem to be true, so that we can think of being true as the concurrence of the different ways of seeming true.

Suppose, then, that this debate turns out in favor of the seems/is distinction for EPI and their properties. Would that reinstate private language? Not by itself. Perhaps two EPI of mine do each possess the feature F, but in order that my word W should apply to them both something else is needed: that W really is my word for F-ness. It is not enough that I do in fact utter W whenever this type of EPI occurs. (I may say "Ouch" whenever a certain type of EPI afflicts me; that does not make "Ouch" *mean* that sort of inner state.) Somehow, most likely to do with my intentions regarding it, I must have given the word W meaning. So: under what conditions can there be meaning, and stable intentions, and does the private linguist fulfill them?[9]

It can certainly be doubted. Saul Kripke (1982) ties the rejection of private language to Wittgenstein's views on following rules, drawing attention in particular to *Philosophical Investigations*, §§201–202 and seeing it as an outcome of the material beginning around §139. Here we are harking back, of course, to the position sketched in §2 of this chapter: meaning requires consistency of use, and consistency can be understood only against the background of communal practice. Therefore no facts about an isolated individual are sufficient to confer meaning; and isn't the speaker of a private language isolated in the relevant sense?

If we take a private language to be one which only its speaker can understand, presumably the answer is yes. But we have seen reason to ignore that part of Wittgenstein's double definition, and think of private language as defined by the epistemic privacy of the objects to which its terms refer. Then the question is less straightforward, since it seems possible (at least until Burke's Assumption can be refuted) that there may be communal linguistic practice relating to EPI; this because Burke's Assumption would allow us to have, and express, beliefs about the private states of others, even though we cannot *know* what these states are like. It will allow us to confirm, from our own impressions, others' statements about their EPI, even to reject their claims on occasion, at least where insincerity or inattention is suspected. So a private language still seems possible; indeed, it looks possible that the public language may also be a private one, in that each of us can use it for referring to their EPI.

Can it be countered at this stage that what we have just described does not count as a communal practice in the required sense, because the required sense demands a practice amongst (so to speak) epistemically equal partners? Without going further into the concept of epistemic equality, it does seem reasonable to agree that Burke's Assumption will not provide it; but will the argument from rule-following really justify the demand? After all, the practice described above is not a mere sham, in which one person makes a statement about their EPI and everyone else respectfully parrots it; the participants are making independent judgments, even if one of them has a favored vantage point. The rule-following arguments are fascinating; they are also complex and controversial, and there must be some room for doubt as to whether they prove the necessity for a communal practice at all. To demand a proof of the need for a community of *epistemically equal* individuals is to impose a substantial further burden on them. Perhaps they can bear it; but that needs to be shown carefully and explicitly.

Two relatively minor points should be mentioned. Christopher Peacocke has proposed what he calls the "Discrimination Principle," claiming that it rules out private language (Peacocke, 1988, especially pp. 491–493). He states it as follows:

> for each content a speaker may judge, there is an adequately individuating account of what makes it the case that he is judging that content rather than any other. (Peacocke, 1988, p. 468)

If, then, p and q are different judgments, there must be something about the act of judging that p which distinguishes it from the act of judging that q. And there are propositions which the supporter of private language will have to claim it possible to judge, which, however, do not satisfy the Discrimination Principle with respect to some other proposition which he is committed to distinguishing from them.

We need not enter into the details of the argument. For whilst this approach may offer a *framework* for discussion of the possibility of a private language, or of many other issues, it cannot by itself settle anything. The Discrimination Principle is nothing but a special case of the trivial truth that if two things are different in respect of a certain property (in this case two judgments in respect of their content), then there is some difference between them relevant to the property in question. And from this nothing can follow as to which particular judgments are legitimate, which spurious, until we add some more substantial premises telling us which factors can be relevant to the content of a judgment. But that falls little short of saying: until we add a theory of meaning. The Discrimination Principle leaves everything still to be contested.

Finally, there is one early approach that can with some confidence be written off: the private-language argument is not, and never really has been, based on skepticism about the memory when exercised about previous EPI. A. J. Ayer (Ayer, 1954) appears to have taken it in this way, and consequently had little difficulty in disposing of it with the counter that reliance on memory was equally necessary for the verification of utterances in the public language. Certain it is that this line has no prospects whatever unless backed up by reasons for thinking memory especially fallible in the private case. And this is not what has been argued – not by Schlick, and not, on any plausible account of his intentions, by Wittgenstein. Their view was not that our memory is especially likely to fail us where the properties of past EPI are in question; it was, rather, that there are no genuine statements or thoughts, and hence nothing for the memory to report, whether truly or falsely, reliably or not.

Notes

1 For further discussion of these arguments, see Chapter 20, REALISM AND ITS OPPOSITIONS, §2.

2 We can see already the need to treat the private-language question separately; for all the above arguments turn on the unknowability of an EPI by *others*, and so clearly do not apply to the case of a private language, where there are no others to be considered.

3 In Craig (1982) I called this "the assumption of uniformity." Readers should take it that only the ugly name has been changed. Edmund Burke wrote (1757, p. 13):

> We do and must suppose, that as the conformation of their organs are nearly, or altogether the same in all men, so the manner of perceiving external objects is in all men the same, or with little difference.

4 Here I just baldly state the next major lemma of the argument. Fuller discussion will be found elsewhere in this volume: see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §2.

5 Given, that is, the assumption that meaningfulness always involves rules, so that whatever is relevant to meaning must be capable of playing the same role consistently.

6 Schlick favors putting all this in terms of a distinction between "Form" and "Content," thus introducing complexities which I have here attempted to skirt round. His main thesis throughout this sequence of lectures is that language can convey only Form, never Content, and this he bases on a view of meaning not wholly unrelated to the notorious "Picture Theory" of Wittgenstein's *Tractatus Logico-Philosophicus*, to the effect that an utterance can only express a fact if it shares its Form. Some may like to read the lectures with this connection in mind.

7 See n. 6 above.

8 It might be thought that the revision cannot be avoided, since only events which had no corresponding outward sign could be epistemically private in the original sense. But this seems wrong; the fact that an inner state has an externally observable correlate means that others have a clue as to when it is occurring, but that is far from saying that the outward sign is so revealing that they can know from it what the inner event is like.

9 In this connection see Wright (1991).

References

Ayer, A. J. 1954. "Could language be invented by a Robinson Crusoe?" *Proceedings of the Aristotelian Society*, suppl. vol. 28: 63–94. Also reprinted in Jones, 1971, pp. 50–61.

Burke, E. 1990 (1757). *A Philosophical Enquiry into the Origin of our Ideas of the Sublime and Beautiful*, edited by A. Phillips. Oxford: Oxford University Press.

Craig, E. J. 1982. "Meaning, use and privacy." *Mind*, 91(364): 541–564.

Dummett, M. 1973. "The philosophical basis of intuitionistic logic." Reprinted in *Truth and Other Enigmas*, pp. 215–247. London: Duckworth.

Harrison, R. 1974. *On What There Must Be*. Oxford: Clarendon Press.

Jones, O. R., ed. 1971. *The Private Language Argument*. London: Macmillan.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

Peacocke, C. 1988. "The limits of intelligibility: a post-verificationist proposal." *The Philosophical Review*, 97(4): 463–496.

Schlick, M. 1938. "Form and content: an introduction, to philosophical thinking." In *Gesammelte Aufsätze*. Vienna: Gerold.

Wittgenstein, L. 1953. *Philosophical Investigations*, translated by G. E. M. Anscombe. Oxford: Blackwell.

Wright, C. J. G. 1991. "Wittgenstein's later philosophy of mind: sensation, privacy and intention." In *Meaning Scepticism*, edited by Klaus Puhl, pp. 126–147. Berlin: Walter de Gruyter.

Further Reading

Blackburn, S. W. 1984. "The individual strikes back." *Synthese*, 58(3): 281–301.
Craig, E. J. 1986. "Privacy and rule-following." In *Language, Mind and Logic*, edited by J. Butterfield, pp. 169–186. Cambridge: Cambridge University Press.
Craig, E. J. 1991. "Advice to philosophers: three new leaves to turn over." *Proceedings of the British Academy*, 76: 265–281.
Wright, C. J. G. 1987. "Introduction." In *Realism, Meaning and Truth*, especially pp. 13–23. Oxford: Blackwell.

Postscript

GUY LONGWORTH

Two central questions about meaning and privacy are the following. First, could there be a private language – a language the expressions of which have meanings that are available in principle to only one person? Second, assuming that the languages that we employ are not entirely private in that broad sense – so that at least some of the meanings they carry are available, in principle, to pluralities of speakers – does that fact serve to rule out otherwise plausible views about the natures of those meanings? In particular, does it rule out views on which meanings determine reference to elements about which only one person can know? The earlier part of this chapter focused largely on the second question, and this part retains that focus. (For discussion of the former question, with reference to Wittgenstein's important discussions, see Wright, 1986, and Stern, 2011.)

I'll focus on a specific range of issues that arise from the conjecture that the meanings of some of an individual's expressions purport to determine reference to objects, conditions, processes, or occurrences that are private to that individual – say, the individual's experiences, or aspects of their experiences. And I'll begin with an initial characterization of what it is for such referents to be private, according to which their privacy entails that they are such that only one person can have propositional knowledge about their existence (or obtaining, occurrence, or unfolding) or their intrinsic properties. (The qualification that the properties be intrinsic is aimed at finessing technical issues that arise from someone knowing, about a referent, only that they know almost nothing about that referent.) We'll later briefly consider a weakening of the operative characterization of privacy. (For discussion of Wittgenstein's animadversions against the claim that experiences, or aspects of experience, are private see Snowdon, 2011.)

A natural and popular view is that understanding an expression is a matter of knowing what it means. More specifically, it might be held that one understands a use of an expression if and only if, for some meaning *M*, one knows that the expression, as so used, means *M*. Suppose in addition that propositional knowledge of the meaning of an expression requires propositional knowledge about the referents determined by the expression. That is, suppose that one knows that an expression means *M* only if, for all referents *R* determined by *M*, one knows that the expression refers to *R*. On that supposition, the natural view about understanding would deliver a straightforward argument from the privacy of referents to the impossibility of more than one person understanding an expression the meaning of which determined those referents. However, there are apparently reasonable grounds for skepticism about the natural view of understanding.

Is it true, then, that if one understands an expression, then there is a meaning *M* such that one knows that the expression means *M*? In the earlier part of the chapter, the claim was challenged on the grounds that reliably true belief about what expressions mean – perhaps as supported by further reliably true belief that the beliefs about meaning are reliably true – might suffice for understanding those expressions. (See the discussion there of what Craig calls Burke's Assumption, pp. 252–257, 263–264.)

Additional grounds for doubt may be provided by reflection on cases in which a speaker appears to understand the use of an expression and yet appears not to meet plausible conditions on knowing what it means. For instance, it's a plausible necessary condition on knowing what an expression means that one has relevant beliefs about what it means – for example, that where one knows that the expression means *M*, one believes that it means *M*. However, consider a case in which you believe that you are in the control of an evil super scientist who has made it so that all of your present experience is hallucinatory. Because of your belief about your circumstances, you withhold belief in how things appear experientially to you. Thus, for example, when you have an experience of a person talking to you, you take it to be the upshot of a hallucination and so do not form the belief that there is someone talking to you. And when you have an experience of the person saying to you that it's a nice day, you fail to form the belief that the person has said to you that it's a nice day. So, you don't believe that the person said to you that it's a nice day. On the assumption that it's impossible for one to know that they said that it's a nice day if one doesn't believe that they said that it's a nice day, this would be a circumstance in which you don't know that they said that it's a nice day. Despite that, it's plausible that – since the person did in fact say that, and you were aware of them doing so – you understood what they said. So, there are plausible grounds for allowing that it's possible to understand what someone said – or to know what, on that occasion, they meant – without knowing what they said. (See Hunter, 1998; Longworth, 2008.)

Another way of generating such doubts would be the following. It's a plausible condition on knowing that such-and-such that, in believing as one does, one couldn't too easily have been wrong. Thus, if, in sufficiently similar circumstances, one would have formed a similar belief, in a sufficiently similar way, and yet that belief would have been false, then it's plausible that, in the actual case, one doesn't know. One wouldn't know because one's true belief could too easily have been false. Given that plausible condition on knowing, grounds for doubt about the claim that if one understands, then one knows, would be provided by plausible cases in which one understood an expression by believing it to mean what it in fact means, and yet one's belief could too easily have been false.

Consider the following case. Imagine that you have acquired competence in English via a normal route. Normally, we can imagine, you are in a position to understand what people are saying to you in English and, moreover, to know what they are saying to you. However, your present circumstances are not normal. You are presently under the control of an evil super scientist, so that the vast majority of your present experiences are hallucinatory. Although the contents of your hallucinatory experiences are prosaic, their natures are such that most of the beliefs that you form naturally on their basis will be false. However, for a few moments each day the scientist relinquishes control of your experience, in such a way that the transition between hallucinatory and genuine experience is not noticeable. During one of these periods, the scientist speaks to you in English, saying to you that it's a nice day. Plausibly, during your brief respite from hallucinations, it is possible for you to see and hear the scientist speaking to you. That is so, despite the fact that it is plausibly impossible for beliefs that you form on the basis of what you see and hear in those circumstances to be

knowledgeable, since it would be so easy for you to have so believed and to have been wrong. Furthermore, and of more immediate relevance, it's plausible that it's possible for you to understand what the scientist says to you. Again, that seems plausible despite the fact that it's plausible that you can't know what the scientist is saying to you since it would be so easy for you to have so believed that he was saying it and to have been wrong. So, it's plausible that it's possible to understand what was said in one's presence even in cases in which one doesn't know what was said. (See Pettit, 2002; Longworth, 2008.)

Suppose that that's right. It would follow that it's possible to meet necessary conditions on understanding what was said without thereby meeting sufficient conditions on knowing what was said. In that case, even if we were to accept that it is not possible for one to know what was said – because, for example, that would require knowledge about the referents of what was said, and that is precluded because of their privacy – it wouldn't follow immediately that it is not possible to understand what was said. For it may be that the reason that knowledge of what was said is precluded is because one's circumstances mean that there is too great a danger that one's beliefs about what was said might be false. Since – by the present supposition – one can understand what was said even in cases in which there is a significant danger of believing falsely, that would present no bar on one's understanding what was said.

In summary, whether the impossibility of knowing what an expression is used to mean translates into the impossibility of understanding what the expression is used to mean depends on two factors. First, it depends on which conditions on knowing figure in determining the impossibility of knowing what the expression means. And, second, it depends on whether those conditions are also conditions on understanding what the expression means.

Let's set that issue to one side and turn to a different range of issues that arise if we assume that understanding what is said requires knowing what is said. The initial characterization of privacy about referents was that a private referent is something such that only one person can have propositional knowledge about its existence or its intrinsic properties. It might seem to follow straightforwardly from that characterization of privacy that it's impossible for more than one person to know what someone says about such a private referent. As a concession to that line of thought, let's assume that someone's saying something about such a referent entails that the referent occurred or that it has one or another intrinsic property. And suppose that a strong closure condition applies to knowing. Specifically, suppose that if one knows that *p*, and if it being so that *p* entails it being so that *q*, then one is in a position to know that *q*. (Compare Craig on p. 253.) Given those assumptions, it's plausible that no one other than the speaker could know what the speaker says about their private referents. For suppose that a speaker *A* says that *p*, and that they are the only person able to know about the existence or the intrinsic properties of a referent determined by the proposition that *p*. And now assume, for purposes of *reductio* that a speaker *B*, distinct from *A*, comes to know that *A* said that *p*. *A*'s having said that *p* entails that the referents determined by the proposition that *p* occurred or that they have one or another intrinsic property. By the strong closure condition, it follows from *B*'s knowing that *A* said that *p* that *B* is in a position to know that the referents occurred or that they have one or another intrinsic property. But by assumption, since at least one such referent is private according to the initial characterization, *B* is not in a position to know that that referent occurred or that it has one or another intrinsic property. And so, given our assumptions, the supposition that *B* knows that *A* said that *p* entails a contradiction and, so, must be rejected.

An obvious problem with that derivation is that it depends upon the correctness of the strong closure condition. For the strong closure condition is obviously too strong, entailing,

for example, that if one knows the Peano–Dedekind axioms, then – regardless of one's knowledge of what those axioms entail – one is thereby in a position to know all arithmetical facts. Furthermore, the natural weakening of the closure condition would be too weak to deliver the required conclusion. For the natural weakening would be the following: if one knows that *p*, and if *one knows* that it being so that *p* entails it being so that *q*, then one is in a position to know that *q*. And that condition would have as a consequence that it is impossible for *B* to know that *A* said that *p* only if *B knew*, of the private referents determined by *A*'s saying that *p*, that *A*'s having said that *p* entails that those private referents occurred or that they had one or another intrinsic property. Not only is there no immediate reason to think that an arbitrary *B must* have the required knowledge but, in addition, there is reason to think that, because of the privacy of the subject-matter of the entailment's consequent, *B can't* have the required knowledge. So, even if the weaker closure condition were acceptable – and even the weaker condition is, at best, controversial – it would not obviously furnish the required connection between the knowledge of meanings and knowledge of referents determined by those meanings. (See, e.g., Dretske, 1970; Hawthorne, 2005.)

Perhaps, however, there is an alternative form of closure condition that is more plausible than the strong closure condition and that, in addition, sustains the required consequence. Although it doesn't appear to be true in general that knowledge is closed under entailment, perhaps it's true that knowledge is closed under a subclass of cases of entailment. Suppose that the subclass could be shown to include all cases in which the entailment relation is underwritten by the connection between someone saying something and the existence or the intrinsic qualities of the referents that are determined by what they say. In that case, closure would apply in the case at issue, connecting knowledge of what *A* said and knowledge about the private referents determined by what *A* said. And that would suffice to reinstate the conclusion that *B* cannot know what *A* said.

Rather than pursue the large and delicate question whether the required form of closure condition is defensible, I propose instead to consider a more local issue. Assume that the required form of local closure condition is correct. We've seen that that closure condition, in conjunction with referents that are private according to our initial characterization, entails that at most one person can understand what is said when what is said determines those referents. For according to the initial characterization of a private referent, it is something such that only one person can have propositional knowledge about its existence or intrinsic properties. The difficulty with that as an argument to our target conclusion is that in assuming that initial characterization of privacy, it seems to assume an overly demanding conception of privacy.

The initial conception seems overly demanding for the following reason. It is liable to seem natural – at least to a proponent of private referents – to hold that there are referents that are available, for instance as objects of knowledge, in a particular way to only one person. For example, a proponent of private referents might hold that there are objects of private experiences that can be known in a peculiarly first-personal way only to their subject. Furthermore, it is arguable that, unlike objects of public experience, such referents can be known at first hand to only one person, the subject of the experience. However, even if such a conception of private referents were defensible, it would not immediately support the claim that such referents are unknowable to anyone other than their subject. In order to support the latter claim, a case would have to be made that such referents are not knowable at *second* hand, on the basis of understanding and accepting what the subject says about them.

The considerations that we've discussed to this point seem powerless to sustain such a case. For example, any such case would have to go beyond appeal to the weak closure condition considered above, on which knowledge is closed under known entailment. For the weak closure condition would serve, at most, to impose the requirement that one who understands what is said about a private referent is placed, thereby, in a position to acquire knowledge about that referent. (At most, since, as was noted above, it would deliver that result only with respect to subjects who knew the operative entailments.) And one might meet that condition via understanding what a subject tells one about the referent, even if no non-vicarious route to such knowledge were available to one. For example, one might gain the required epistemic position by reasoning from facts about meaning and facts about what those facts entail. What would be needed, in addition, in order to make the required case would be a defense of the further condition that one must be in a position to know the known entailments of what one knows independently, and in advance, of knowing the entailing facts. Furthermore, our ordinary judgments about cases seem consistent with allowing the possibility of purely second-hand knowledge about the things to which an interlocutor refers. For example, it seems possible for you to acquire knowledge about someone with whom you are not acquainted – say, my friend Kim – solely on the basis of my testimony about them. Argument would be required to show that one couldn't similarly come to know about another's private referents on the basis of their testimony. (See, e.g., Evans, 1982, pp. 122–129; 1985, pp. 6–8; Raven, 2008.)

Let me summarize the foregoing. I've considered the prospects for an argument from the privacy of a referent determined by the meaning of an expression to the impossibility of more than one person understanding a use of the expression. I considered a line of argument based on an initial characterization of privacy, according to which a private referent is something such that only one person can have propositional knowledge about its existence or intrinsic properties. The argument was, in effect, based upon two assumptions: first, that understanding the meaning of a use of an expression requires knowing the meaning of the expression; second, that knowing the meaning of an expression requires knowing about the referents of the expression. I suggested some grounds for doubting both assumptions. Finally, I suggested that the initial characterization of privacy might be too demanding. And the adoption of a more reasonable characterization of privacy would further dim the prospects for a compelling argument from privacy of referents to the impossibility of mutual understanding.

References

Dretske, F. 1970. "Epistemic operators." *Journal of Philosophy*, 67(24): 1007–1023.

Evans, G. 1982. *The Varieties of Reference*, edited by J. McDowell. Oxford: Clarendon Press.

Evans, G. 1985. *Collected Papers*. Oxford: Clarendon Press.

Hawthorne, J. 2005. "The case for closure." In *Contemporary Debates in Epistemology*, edited by M. Steup and E. Sosa, pp. 26–42. Oxford: Blackwell.

Hunter, D. 1998. "Belief and understanding." *Philosophy and Phenomenological Research*, 53(3): 559–580.

Longworth, G. 2008. "Linguistic understanding and knowledge." *Noûs*, 42(1): 50–79.

Pettit, D. 2002. "Why knowledge is unnecessary for understanding language." *Mind*, 111(3): 519–550.

Raven, M. J. 2008. "Problems for testimonial acquaintance." *Noûs*, 42(4): 727–745.

Snowdon, P. 2011. "Private experience and sense data." In *The Oxford Handbook of Wittgenstein*, edited by O. Kuusela and M. McGinn, pp. 402–428. Oxford: Oxford University Press.

Stern, D. 2011. "Private language." In *The Oxford Handbook of Wittgenstein*, edited by O. Kuusela and M. McGinn, pp. 333–350. Oxford: Oxford University Press.

Wright, C. 1986. "Does *Philosophical Investigations* I. 258–60 suggest a cogent argument against private language?" In *Subject, Thought, and Context*, edited by J. McDowell and P. Pettit, pp. 209–266. Oxford: Clarendon Press.

12

Tacit Knowledge

ALEXANDER MILLER

1 Introduction

Competent speakers of a natural language know what the sentences of that language mean. A theory of meaning for a natural language, if correct, specifies what each well-formed declarative sentence of that language means.[1] Thus, the following question naturally suggests itself: What sort of relationship, if any, obtains between speakers of, and a correct theory of meaning for, a given natural language? In this chapter I shall examine a number of answers that have been given in response to this question. In particular, I shall be considering whether any account of the relationship can provide an adequate justification for what has been an article of faith of those engaged in the construction of systematic theories of meaning for natural languages: the requirement that such theories be compositional. A theory of meaning is compositional if and only if (a) it has only finitely many proper (non-logical) axioms, and (b) each of the meaning-delivering theorems ("meaning-specifications") served up is generated from the axiomatic base in such a way that the semantic structure of the sentence concerned is thereby exhibited.[2]

What motivation is there for seeking compositional semantic theories in preference to their more readily available, non-compositional counterparts? Why should the construction of a semantic theory be constrained by the requirement that it reflect the semantic structure of the language concerned? As Crispin Wright notes, in a wide-ranging survey of these issues (1986; see also his 1980, ch. 15; 1981; 1988), the answer generally given to this question is that the construction of such theories is supposed to take us some way towards providing answers to each of the following three questions:

1. How is it possible, given the finitude of their capacities, for speakers of a natural language to understand a potential infinity of sentences?
2. How is it possible to *learn* a natural language?
3. How is it possible to understand utterances of previously unencountered sentences?

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

In what follows I shall ignore (1), and look only at (2) and (3). It is not clear to me what the claim that speakers understand a potentially infinite number of sentences amounts to and, in any case, as we shall see, Gareth Evans makes it clear that the demand for compositionality in semantic theories has nothing essentially to do with this alleged "potential infinity": it can be leveled with equal force at theories dealing with languages containing only a finite number of possible sentences. I shall proceed as follows. I begin, in §2, with Michael Dummett's idea that answers to (2) and (3) might be facilitated if competent speakers of a natural language can be credited with *tacit knowledge* of the axiomatic base of a compositional theory of meaning for their language. I shall then explain that viewing tacit knowledge of semantic axioms as a bona fide propositional attitude-state is implausible, because a plausible constraint (outlined by Evans and Wright) on a state's being a propositional attitude is thereby violated. In §3 I examine Gareth Evans's suggestion that ascription of tacit knowledge of a semantic theory can be empirically well founded, so long as we are clear that in ascribing it we are not ascribing a set of genuine propositional attitudes but only a set of *mere dispositions*, one for each primitive expression of the language, to the speaker. Crispin Wright has raised a number of objections against Evans's account, and in §4 I shall show that Evans's suggestion, as developed and modified by Martin Davies, has the resources to respond to those objections. §5 briefly looks at how the modified account can provide answers to questions (2) and (3) above. In §6 I argue that Wright's alternative to Davies's mirror constraint actually *presupposes* it. I finish, in §7, by considering whether the project of constructing semantic theories in accordance with the mirror constraint is in tension with Wittgenstein's reflections on rule-following.

2 Tacit Knowledge and Propositional Attitudes

Dummett writes:

> A theory of meaning will, then, represent the practical ability possessed by a speaker as consisting in his grasp of a set of propositions: since the speaker derives his understanding of a sentence from the meanings of its component words, these propositions will most naturally form a deductively connected system. The knowledge of these propositions that is attributed to a speaker can only be an implicit knowledge. In general, it cannot be demanded of someone who has any given practical ability that he have more than an implicit knowledge of those propositions by means of which we give a theoretical representation of that ability. (Dummett, 1976, p. 70)

It is clear from this passage that tacit knowledge of a semantic theory's axiomatic base is taken by Dummett to be a species of knowledge: that the state of tacitly knowing an axiom of such a theory is taken to be a propositional attitude-state, a state which represents the information codified in that axiom. As Wright comments,

> The explanatory ambitions of a theory of meaning would seem to be entirely dependent upon the permissibility of thinking of speakers of its object language as knowing the propositions which its axioms codify and of their deriving their understanding of (novel) sentences in a manner mirrored by the derivation, in the theory, of the appropriate theorems. (Wright, 1986, p. 207)

It is *tacit* knowledge because competent speakers will generally be unable to formulate the theory of meaning whose axiomatic base they tacitly know, and will generally be unable to recognize a correct formulation of the theory of meaning if it is presented to them. But for all

that, it is still knowledge, and knowledge of propositions. It is easy to see how this contributes to answering (2) and (3) above. Say that a language L is learnable when it is possible for its speakers "to come to know the meanings of all the sentences of L by way of exposure and projection" (Davies, 1981a, p. 60). Thus, we can say that a language is learnable when one needs explicit training with only a relatively small number of sentences in order to secure competence with a possibly very large set of sentences outwith that set. So we can see that to say that a language is learnable is just to say that speakers can understand novel utterances, without explicit training in their use. Question (2) collapses, therefore, for natural languages at any rate, into (3). And there is no problem for Dummett in answering (3): speakers can understand novel utterances because they have at their disposal the information, codified in their tacit knowledge of the theory of meaning for the relevant language, of the semantic properties of the sentence's parts. This tacit knowledge provides them with the resources for understanding the sentence, in the same way that the axioms of the theory of meaning provide the resources for the derivation of a meaning-specification for the relevant sentence.

There is, no doubt, much to be said about Dummett's idea. But for the rest of this section I want to focus on a set of arguments whose upshot is that, whatever tacit knowledge of the axiomatic base of a semantic theory is, it cannot be construed as a genuine propositional attitude or intentional state.

Evans and Wright have argued that there is a necessary condition which all genuine intentional states must satisfy to qualify as such, and that putative states of tacit knowledge of meaning-theoretic axioms do not satisfy this condition. In order to motivate this condition they ask us to contrast, on the one hand, the belief that a man might have to the effect that a certain substance is poisonous, with the disposition that a rat might have to avoid a similarly contaminated substance. Can we describe the rat as having a genuine belief that the substance is poisonous? Evans and Wright suggest not: for whereas in the case of the man the belief is, to use Evans's phrase, "at the service of many distinct projects," and can interact with others of his beliefs and desires to produce new beliefs and desires, none of this obtains in the case of the rat's disposition. In the case of the man, for instance, the belief could be at the service of projects such as killing an adversary, retaining good health, or getting out of an obligation by taking a small dose, to name but a few. None of this is possible in the case of the rat: the putative "belief" is harnessed to the single "project" of avoidance of the substance. This is supposed to be a reflection of the fact that propositional attitudes and intentional states, such as beliefs and desires, come in *articulated systems* or *holistic networks*. And it is because genuine beliefs come in such networks and can thus interact with other beliefs, that they can indeed be at the service of many distinct projects. For example, the man's belief that the substance is poisonous can be at the service of the project of getting out of a particular obligation because that belief can, together with the beliefs that a small amount of the substance causes only a mild illness and that a mild illness will release him from the obligation, lead to the belief that taking a small amount of the substance will enable him to avoid fulfilling the obligation.

A crude version of the constraint suggested by the Evans–Wright discussion might therefore run as follows:

> *Constraint 1*: A state P of an agent W is a genuine propositional attitude or intentional state only if P can interact with others of W's propositional attitudes and intentional states to produce new propositional attitudes or actions – thus putting P at the service of many distinct projects of the agent.[3]

Where does this leave the tacit knowledge a speaker might have of a meaning-theoretic axiom? Evans and Wright both claim that such a state of a speaker violates the constraint above. Far from being at the service of many distinct projects, the tacit knowledge is, says Evans,

> exclusively manifested in speaking and understanding a language; the information is not even potentially at the service of any other project of the agent, nor can it interact with any other beliefs of the agent (whether genuine beliefs or other "tacit" beliefs) to yield further beliefs (Evans, 1981, p. 133)

while Wright puts it like this:

> The (implicit) knowledge of a meaning-theoretic *axiom* would seem to be harnessed to the single project of forming beliefs about the content of sentences which contain the expression, or exemplify the mode of construction, which it concerns.

He asks the following (rhetorical) questions:

> What is supposed to be the role of *desire*? What is the (implicit?) desire which explains why the subject puts his axiomatic beliefs to just this use, and what are the different uses to which they might be put if his desires were different? (Wright, 1986, pp. 227–228)

No plausible answers to these questions suggest themselves, so the conclusion is that states of tacit knowledge of semantic axioms cannot plausibly be viewed as propositional attitudes. Evans appears to view the objection as applying to tacit knowledge *tout court*, that is, not only to states of tacit knowledge of axioms but also to states of tacit knowledge of meaning-theoretic *theorems*, which codify the rules governing the use of whole sentences. But Wright quite clearly sees the objection as applying only to tacit knowledge of the axioms; as he says, "someone who is credited with implicit knowledge of a meaning-delivering theorem may express his knowledge in an indefinite variety of ways, including, in appropriate contexts, lying, assent, and silence." So no reason emerges "to doubt the propriety of crediting [speakers] with implicit knowledge of the content of meaning-delivering theorems" (Wright, 1986, pp. 227 and 237–238). Thus, Wright's defense of genuinely intentional tacit knowledge of meaning-specifying theorems is in effect a claim that tacit knowledge of such theorems can be inferentially integrated with the rest of the agent's propositional attitudes in the manner required by Constraint 1.[4] But how can my semantic belief concerning the meaning of a given sentence interact with my propositional attitudes to give rise to new propositional attitudes? An example should suffice to convince us that this is indeed possible. I have a certain intentional state, the possession of which is constitutive of my understanding of the sentence "Catriona is getting married to Seamus on Saturday": this intentional state can interact with my belief that Catriona is getting married to Sean on Saturday to lead to the belief that the sentence "Catriona is getting married to Seamus on Saturday" is false; or it can interact with my desire to annoy Patrick (an unsuccessful suitor of Catriona's) to lead to the belief that I ought to utter "Catriona is getting married to Seamus" in Patrick's presence; and so on. Examples can quite easily be multiplied: this shows how the tacit knowledge of semantic theorems, unlike the tacit knowledge of semantic axioms, can indeed be at the service of many distinct projects of the agent concerned.

Let's suppose, as seems plausible, that Wright is correct in claiming that the objection just considered does not apply to states of tacit knowledge of the meaning-specifying

theorems, for the reason stated. Then the following question suggests itself: Why cannot we view the tacit knowledge of an axiom as a genuine intentional state after all, in virtue of the fact that although it is *directly* harnessed to the single project of forming beliefs about the content of sentences in which it figures, it can make an *indirect* contribution to the other projects of the agent *via* the states of tacit knowledge of the theorems corresponding to those sentences? The state of tacit knowledge of an axiom is at the service of many of the agent's projects because the project to which it is *directly* harnessed is itself at their service. Or equivalently, a state of "tacit knowledge" of an axiom can lead inferentially to a vast number of other propositional attitudes because it *can* lead inferentially to genuine intentional states (i.e., those that consist in implicit knowledge of the appropriate theorems), which *in turn* can lead to almost any other intentional state, modulo the other intentional states which we suppose the agent to possess.

So what is important in determining whether a state is a genuine propositional attitude is not the *number* of intentional states to which it can give rise, or the *number* of the agent's projects which it is at the service of; rather, what is crucial is the nature of the potential *routes* from the given state to the rest of the propositional attitudes, and the nature of the potential routes via which the information in question is placed at the service of a multiplicity of the agent's projects. I suggest, then, that in order to draw the required distinction in such a way that tacit knowledge of semantic axioms is excluded, we need something along the lines of the following amended version of Constraint 1:

> *Constraint 2*: A state P of an agent W is a genuine propositional attitude or intentional state *only if* P can interact *directly* with others of W's propositional attitudes and intentional states to produce new propositional attitudes and actions – thus putting P *immediately* at the service of many distinct projects of the agent.

Of course, the questions facing us now are: (a) What exactly do we mean by "directly" as it appears in Constraint 2? and (b) Does Constraint 2 provide us with a plausible means of drawing the distinction between intentional and non-intentional states?

Let's attempt to answer (a) first. We can say that a cognitive state P interacts directly with a given propositional attitude only if the interaction takes place without the mediation of some other propositional attitude R in the causal generation of which P plays a part. Getting clear on why states of tacit knowledge of axioms fail to satisfy the constraint will help secure our grip on this notion of directness. Such a state can interact with other states only via the states of tacit knowledge concerning the sentences in which the expression corresponding to the axiom figures, because of Frege's insight that a speaker's understanding of a subsentential expression can be manifested only through the use that he makes of whole sentences in which that expression figures: it is always by means of complete sentences that we perform linguistic acts or, more figuratively, make moves in a given language game. Thus, suppose that there is a cognitive state of mine which represents the information that a given predicate, for example, "horse," has such-and-such satisfaction conditions. Suppose also that I hear someone utter the sentence "I have a horse with five legs." Then I might form the belief either that the person in question simply has an understanding of the predicate which differs from mine, or that he has a very rare and unusual sort of horse. But these beliefs can be formed *only* via my implicit knowledge of the meaning of the whole sentence "I have a horse with five legs," because it is only in the context of a whole sentence that a linguistic act involving the predicate can be effected.

So, tacit knowledge of an axiom can never interact directly with other putative intentional states, because it always has to interact with them via states of tacit knowledge of the appropriate meaning-theoretic theorems, in whose causal generation it plays a part.

But now for question (b): Is the fact that a given state fails to satisfy Constraint 2 good grounds for refusing to describe that state as genuinely intentional? Two further questions we might ask in attempting to decide on the plausibility of Constraint 2 are: (1) Is there any good *a priori* motivation for the constraint?, and (2) Does the constraint rule out the states which intuitively ought to be ruled out?

I won't spend a great deal of time on (2). I will limit myself to noting, first, that no genuine belief state can be ruled out by Constraint 2 since I can move from the belief that P to almost any other belief that Q quite simply, by coming to possess the belief that $P \rightarrow Q$ and drawing out the appropriate inference: where the interaction between the belief that P and the belief that $P \rightarrow Q$ needn't take place via any further intentional state causally generated by the belief that P.[5] And, second, that the constraint does appear to rule out at least some of Stich's intuitive examples of subdoxastic states: in the case of Hess's experiment with the retouched photographs, for example, it seems that it is only via their role in generating genuine beliefs that the states representing information about pupil size can interact with other of the agents' propositional attitudes.[6]

So I would tentatively suggest that we can give an affirmative answer to (2). But what about (1)? Why should a state which fails to satisfy Constraint 2 be discounted from being a genuine intentional state? I think this latter question can only be answered after some reflection on the role played by the postulation of intentional states in the rationalistic explanation of human behavior. Some sorts of behavior exhibited by a human agent call for explanation in terms of the beliefs, desires, and other propositional attitudes possessed by that agent, and the use of language is clearly one such kind of behavior. So we attempt rationalistically to explain a person's use of his language by crediting him with a range of intentional states. The crucial point is that once we have credited him with intentional states corresponding to the theorems of a correct theory of meaning for his language, we have everything we need in order to run the appropriate explanation: crediting the speaker with intentional states corresponding to the *axioms* adds nothing whatsoever to the rationalistic explanation of the speaker's behavior provided by ascribing to him intentional states concerning the rules for the use of whole sentences. The explanatory redundancy of the ascription of states of tacit knowledge of the axioms is guaranteed by the fact that they only ever play a part in explaining behavior via states of tacit knowledge of theorems: if it were possible for states of tacit knowledge of axioms to interact directly with other intentional states, then this crucial point about explanatory redundancy could not be made.

So it seems that room can be found for the states corresponding to semantic axioms only within a purely causal explanation of speakers' behavior, and accordingly we can view such states only as purely causal states which play a part in the proximate causal history of the (intentional) states corresponding to the theorems in the semantic theory. Ascribing to speakers intentional states corresponding to various parts of the axiomatic base would simply be to load our explanatory theory with more baggage than is warranted by its explanatory brief: if P only ever "interacts" with other propositional attitudes via a state P* of which it is a causal antecedent, and if describing P as intentional exceeds the explanatory demands on the theory – in the sense that describing P as an intentional state makes *no* contribution whatsoever to that explanation – then it seems that the most that we can claim concerning P is that it is a causal state which plays a part in the proximal causal history of P*.[7]

This seems to rule out Dummett's idea that states of tacit knowledge of the axioms of a semantic theory can be viewed as genuine propositional attitudes. Is there an alternative way of construing the relationship between speakers and the axiomatic base of the semantic theory, and can we still justify the demand for compositionality? Evans's alternative to Dummett attempts to answer these questions, so it is to Evans's discussion that we now turn.

3 Tacit Knowledge and Dispositional States

Evans's discussion proceeds with reference to the relatively simple, and finite, language L consisting of 10 names "a," "b," "c," …, which stand for Harry, John, Bill, …, and 10 predicate expressions "F," "G," "H," …, which stand for happiness, baldness, heaviness,.... L thus has 100 syntactically admissible sentences, each consisting of the concatenation of a name with a predicate.

Suppose that a semantic theorist sets out to find the correct theory of meaning for this language. One constraint is that the theory settled for should have the right *output*: the meaning-specifying theorems which it issues in should be *correct*. Suppose that what the semantic theorist is after is a correct, Davidson-style truth-conditions theory for L. Then we will regard the theory as acceptable if it delivers the following set of truth-conditions specifications:

"Fa" is true-in-L iff Harry is bald
"Fb" is true-in-L iff John is bald
…
"Ga" is true-in-L iff Harry is happy
…
"Oj" is true-in-L iff Michael is anxious.

But then the following problem arises. Call two theories which issue in the same set of truth-conditions specifications *extensionally equivalent*. Then the following two theories for L will be extensionally equivalent:

T1: the *listiform* theory, which has 100 axioms, one for each individual sentence of L (i.e., simply the full list of truth-conditions specifications given immediately above).

T2: the articulated theory consisting of 21 axioms, one for each of the proper names (e.g., "a" denotes Harry), one for each of the predicates (e.g., an object satisfies "F" iff it is bald), and an axiom for the subject-predicate mode of combination (e.g., "a sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate").

T2 is clearly closer in spirit than T1 to the theories which semanticists have in fact been attempting to construct for natural languages; but Wright's demand for a motivation for compositionality can now be stated as follows: given that the constraint that a theory issue in the correct truth-conditions specifications is not by itself sufficiently strong to discriminate in favor of T2, can any further constraints be imposed which will provide a motive for

the preference of T2 to T1? In other words, can there ever be empirically respectable evidence which will discriminate between two extensionally equivalent theories?[8]

Evans suggests the following constraint: the theory should aspire to provide not just the correct truth-conditions specifications, but also a description of the *dispositions* corresponding to each of the expressions for which that theory has a proper axiom. So if we find that speakers have 100 dispositions of the relevant type we will be justified in opting for T1, whereas if we find that they have 20 such dispositions, the acceptance of T2 will be warranted. But what are the dispositions "of the relevant type" alluded to above? In the case of T1, the dispositions corresponding to its primitive expressions (the expressions to which it devotes an individual axiom) are easy to specify: each disposition is simply "a disposition to judge utterances of the relevant sentence-type as having such-and-such truth-conditions" (Evans, 1981, p. 124). T2 is more problematic because its primitive expressions are not whole sentences, but rather proper names and predicate expressions, and it is only sentences which can be said to have truth-conditions. As a consequence, the dispositions corresponding to the primitive expressions of T2 have to be interdefined. Evans suggests characterizing the dispositions as follows:

> We might say that a speaker U tacitly knows that the denotation of "a" is Harry iff he has a disposition such that:
>
> (Πφ)(Πψ) if:
> (i) U tacitly knows that an object satisfies φ iff it is ψ
> (ii) [if] U hears an utterance having the form φ^a, then U will judge that the utterance is true iff Harry is ψ.
>
> Connectedly, we say that a speaker tacitly knows that an object satisfies "F" iff it is bald iff he has a disposition such that:
>
> (Πx)(Πα) if:
> (i) U tacitly knows that the denotation of α is x
> (ii) [if] U hears an utterance having the form F^α, then U will judge that the utterance is true iff x is bald.
>
> In these formulations, "Π" is a universal substitutional quantifier, with variables having the following substitution classes: φ, names of predicate expressions of the (object) language, α, names of names of the (object) language: ψ, predicate expressions of our language (the metalanguage), and "x," proper names of our language. (Evans, 1981, pp. 124–125)

How can we tell whether or not a speaker has the dispositions possession of which constitutes tacit knowledge of T1 or of T2? Evans suggests three sources of empirical evidence.

The first source is connected with Evans's insistence that the notion of a disposition involved in his account has to be taken in a full-blooded way: the ascription of a disposition is not to be regarded merely as a statement that some regularity obtains:

> These statements of tacit knowledge must not be regarded as simple statements of regularity, for if they were, anyone who correctly judged the meanings of complete sentences would have a tacit knowledge of T2. When we ascribe to something the disposition to V in circumstances

> C, we are claiming that there is a state S which, when taken together with C, provides a causal explanation of all the subjects V-ing (in C). So we make the claim that there is a common explanation to all these episodes of V-ing. Understood in this way, the ascription of tacit knowledge of T2 … involves the claim that there is a single state of the subject which figures in a causal explanation of why he reacts in a regular way to all the sentences containing the expression …. The decisive way to decide which [ascription of tacit knowledge] is correct is by providing a causal, presumably neurophysiologically based, explanation of comprehension. With such an explanation in hand, we can simply see whether or not there is an appeal to a common state or structure in the explanation of the subject's comprehension of each of the sentences containing the proper name "a." (Evans, 1981, pp. 125–127)

In addition, we can also examine the patterns of acquisition of knowledge of the meanings of sentences manifested in the linguistic behavior of L-speakers. For example, evidence suggestive of tacit knowledge of T1 would be that even when a speaker has acquired dispositions to judge correctly of the truth-conditions of Ga and Fc, he is not thereby (in the absence of further training and exposure) disposed to judge correctly of the truth-conditions of Gc. Evidence suggestive of tacit knowledge of T2 would be that he is, under the same conditions, so disposed.[9]

Further evidence is provided by the patterns of loss of knowledge of meanings exhibited in speakers of L. If such a speaker is initially competent with each of the 100 sentences of L and if, by knocking out his competence with, say, Hd, we thereby disturb his competencies with all other sentences containing the expressions "H" and "d," then tacit knowledge of T2 will be ascribable. But if the other competencies remain undisturbed by the speaker's loss of competence with Hd, then the ascription of tacit knowledge of T1 will be in order.

It seems, then, that we have found an empirically respectable way of deciding which of T1 and T2 should be accepted for the language L. It is perhaps worth noting that at this stage the central idea underlying Evans's account seems to be the following: the derivational structure of a theory of meaning (the canonical routes from its axioms to its theorems) should in some sense reflect the causal structure found among the competencies of the speakers of the language under scrutiny (the causal routes leading from the dispositions associated with the language's names and predicates to the intentional states associated with the whole sentences of the language). In the remainder of this chapter I shall elaborate upon and question the constraint motivated by this central idea.

4 Wright's Attack on Evans

In this section I outline three criticisms that Wright has raised against Evans's dispositionalist account of tacit knowledge of semantic axioms, and the responses that have been offered by Martin Davies on Evans's behalf. I shall argue that the responses offered by Davies to two of Wright's criticisms are unsuccessful as they stand, and sketch my own alternative defense of Evans against them. I will then show how Davies's response to the third objection is plausible as it stands.

Wright's first objection to Evans concerns characterization of the dispositions which constitute tacit knowledge of T2. We can see from the quotations above that the dispositions corresponding to the names and predicate expressions of L have to be interdefined, that is, the dispositions which constitute tacit knowledge of the denotation conditions of the

proper names of the language are defined in terms of tacit knowledge of the satisfaction conditions of L's predicate expressions; and the dispositions which constitute tacit knowledge of the satisfaction conditions of the predicates are defined in terms of tacit knowledge of the denotation conditions of L's proper names. Why is this a problem? As Wright puts it, "to characterize a disposition ought to be to characterize both what it is a disposition to do and the circumstances under which it will be manifest" (Wright, 1986, p. 233). Suppose, for example, that we are trying to give a dispositional account of ductility, and that we come up with: X is ductile iff the observable phenomena $c_1, \ldots, c_n$ occur under background circumstances C. Suppose further that the conditions C include the possession by X of the additional dispositions $d_1, \ldots, d_k$. Wright's point is that if, in characterizing the manifestations distinctive of some one of the further dispositions d_i, say, we have to refer to background circumstances *which include the assumption that X is ductile*, we will have said thereby *nothing whatsoever* as to what ductility consists in: we will simply have failed to say what ductility is.

This point seems to me to be fundamentally correct, and it is easy to see how it applies to Evans's account of tacit knowledge of T2. Take the disposition which constitutes tacit knowledge of the denotation of the name "a." What we are after in characterizing this disposition is something of the form: X tacitly knows that the denotation of "a" is Harry iff observable phenomena $c_1, \ldots, c_n$ occur under background conditions C. In this case the background conditions include the possession by X of the further dispositions $d_i, \ldots$ = tacit knowledge of the satisfaction conditions of certain of the predicate expressions of L. But a characterization of the distinctive manifestations of the d_i's is possible only if we make reference to background conditions in which X is assumed to have tacit knowledge of the denotation conditions of the names of L, and it is precisely *this* species of tacit knowledge which we are trying to explicate. So our account turns out to be viciously circular, and we fail altogether in our attempt to say what tacit knowledge that the denotation of "a" is Harry consists in.

Wright himself is not pessimistic about the possibility of a solution to this problem:

> I offer the point more as something which someone who wished to advance Evans's account should say something about than as an objection. Perhaps a more sophisticated account of the notion of a disposition would remove the worry; my own suggestion would be that Evans's proposal should have proceeded by reference to states of a different sort – his real interest, after all, is in the underlying 'categorical' bases. (Wright, 1986, p. 233)

And Davies subsequently offers what seems to be a respectable way around the trouble threatened by Evans's characterization. Davies suggests that we cast our account of tacit knowledge in terms of "underlying explanatory states," rather than in terms of dispositions; instead of defining tacit knowledge in terms of the dispositions a speaker has concerning truth, satisfaction and denotation conditions, we define it in terms of the states which make up the "categorical bases" underlying those dispositions.[10] The speaker with tacit knowledge of T2 will not now be characterized as having 20 dispositions defined *à la* Evans, but as being the bearer of 20 *causal explanatory states*, each of which is the basis of one of those dispositions. This allows us to state the constraint that was breaking through the clouds at the end of §3:

> If, and only if, a speaker who has dispositions to judge correctly of the truth-conditions of $S_1, \ldots, S_n$ is thereby (and without any further training or exposure) disposed to judge

> correctly of the truth-conditions of S, should the semantic resources sufficient for the canonical derivation of truth-conditions specifications for S_1, …, S_n be sufficient for the canonical derivation of a truth-condition specification for S.

Under our first revision of Evans's account in terms of underlying states, this becomes what Davies terms the *mirror constraint*:

> If, and only if, the operative states implicated in the causal explanation of a speaker's beliefs about the meanings of S_1, …, S_n are jointly sufficient for a causal explanation of his belief about the meaning of S, should the semantic resources sufficient for a canonical derivation of truth-conditions specifications for S_1, …, S_n be sufficient for the canonical derivation of a truth-condition specification for S.

However, I have the following worry about whether this does satisfactorily avoid Wright's problem concerning interdefinability and vicious circularity: If our only means of *individuating* the categorical bases underlying the dispositions is via the dispositions which they underlie, then doesn't the problem simply carry over into the revised account in terms of causally operative states? Is it possible to characterize the causally operative state which underlies my disposition concerned with the denotation condition of the name "a" *without* referring to the causal states underlying the disposition I have connected with the satisfaction conditions of L's predicates? If not, and if this holds vice versa, then I suggest that we have again failed to say what tacit knowledge that the denotation of "a" is Harry consists in, because we will have failed to individuate the causal state, possession of which allegedly constitutes that tacit knowledge. So we are faced with the following dilemma: either we must provide an account of how the operative states can be individuated without reference to the dispositions which it is claimed they underlie – which account is at present lacking – or, on the other hand, the problem that arose for Evans's account arises again for Davies's proposed revision.

Is Wright's objection, then, fatal to Evans's account? We would be over-hasty in concluding that it is: Davies's switch to talk of underlying causal states and categorical bases, indeed, does nothing to remove the circularity which Wright focuses on, but how *vicious* is that circularity? I will suggest that the circularity here will infect any constitutive account of the mastery of individual sub-sentential expressions, no matter whether that account is couched in terms of dispositions, underlying causal states, or anything else we care to choose. This should raise our suspicions about whether circularity can be regarded as a *defect* in such an account.

Any account of what competence with a name consists in will have to contain resources sufficient for an account of what competence with the sentences containing that name consists in, since it is only in the use of whole sentences that competence with the name can be manifested. This means that we will also require an account of what competence with predicate expressions consists in; and when we try to give this latter account – an account of what understanding the sentences containing the predicates consists in – we find ourselves back at the point we started out from, requiring an account of what competence with names consists in. Thus, *any* account of what competence with a name consists in requires an account of what competence with predicate expressions consists in, and vice versa.

This points to the proper line of response to Wright's objection. Isn't it the case that something analogous to this interdefinability property is possessed by beliefs and desires?

Intentional action requires both beliefs and desires to be present, and, more generally, beliefs, desires, and propositional attitudes are ascribable to an agent only in systems, and not individually. When we attempt to give, say, a constitutive account of the belief which partially rationalizes a certain action, we stand in need of a similar account of the appropriate desire, and vice versa. But we would not take this to signal the impossibility of providing a constitutive account of either belief or desire; rather, the conclusion drawn is that the relationship between the beliefs, desires, and the behavioral facts which ground their ascription is irreducibly *holistic* (see Chapter 15, HOLISM). How, then, do we say what beliefs and desires are? I think, roughly speaking, that there are two components to this. First, by showing how the propositional attitude ascriptions relate *to each other* by giving the *a priori* principles which constrain the relations *between* the various sorts of propositional attitude; and second, by giving the interpretative principles which link the propositional attitudes holistically to the behavioral facts which ground their ascription (see Fricker, 1981). In summary, we individuate a propositional attitude not by picking it out individually, but by giving its place in the network of propositional attitudes which form part of any agent's mental armory. So the fact that we *can't* pick out such states in isolation from the entire network in which they occur needn't give us too much cause for concern.

I suggest that an analogous point is available in the case of states of tacit knowledge. All the interdefinability focused on by Wright shows is that tacit knowledge, too, has to be ascribed in a holistic fashion. We do not give an account of what tacit knowledge of, say, a name "a" consists in, apart from an account of what constitutes tacit knowledge of the complete axiomatic base: and we give such an account by delineating the constraints which govern the ascription of tacit knowledge of the axiomatic base *as a whole*. This is where the mirror constraint has a crucial part to play: faced with a semantic theory T, we decide whether or not a speaker should be ascribed tacit knowledge of T by seeing whether the theory meets the mirror constraint with respect to that speaker; in other words, by seeing whether the derivational structure *of the whole theory* is isomorphic in the relevant sense to the causal structure found in that speaker's overall competence. Just as we give an account of what beliefs are by giving their location within a wider network of propositional attitudes, we give an account of what tacit knowledge is by showing how the causal states in question are located in a wider causal structure: having given the structure, we need say nothing more about the composition of the individual states. The mental, unlike the metallurgical, is essentially holistic.[11]

I now move on to look at Wright's second objection to Evans's account of tacit knowledge. I will argue that although the solution which Davies proposes is a good solution to a problem which Evans perhaps ought to have taken account of, it simply fails altogether to engage with Wright's objection in the deeper form in which he originally raised it. I will then show how Davies ought to have responded to Wright's deeper objection.

Davies summarizes the objection thus:

> The second objection relates to the account of tacit knowledge as a certain kind of causal structure. Suppose that a subject knows (tacitly or in the ordinary sense) what the various sentences of L mean; and suppose that underlying those pieces of knowledge there is indeed a causal structure of the kind which, on Evans's account, is required for tacit knowledge of T2 – the articulated theory. Wright asks why such a subject would not be at least as well described by a two-part theory. The first part would be the *semantic* theory T1 – the listiform theory; the second part would be 'some appropriate hypotheses of a *non-semantic* sort, about the presumed

> causal substructure of the dispositions which T1 describes'. Why, in short, does mere *causal* structure justify articulation in a *semantic* theory? (Davies, 1987, pp. 443–444).

The suggested problem seems to be that not all attributions of causal explanatory structure will be pertinent to the ascription of tacit knowledge to the speaker concerned; and in order to bring this point home Davies provides an example of a case in which "a pattern of breakdown is intuitively misleading as to the attributability of tacit knowledge … [but where] … the evidence is not obviously misleading as to the presence of some kind of causal structure" (Davies, 1987, p. 451).

Consider the speaker C who has a language system which performs derivations in an explicit representation of the semantic theory T1 – the matrix in Figure 12.1 is supposed to show how the representations of the 100 axioms of T1 are arranged, and each of its elements represents an information storage unit. Now we elaborate the example somewhat and suppose that in order to function properly the individual units of the matrix have to be supplied with certain nutrients, which flow through the matrix in channels. Suppose also that there are two types of nutrient X and Y, that the X nutrient flows through the matrix in 10 channels X_a, …, X_j "each of which serves the ten storage units for the ten sentences containing a single name," and that in a similar fashion the Y nutrient flows in 10 channels Y_F, …, Y_O "each of which serves the ten storage units for the ten sentences containing a single predicate." The supposition crucial for the example is that each of the storage units will *fail* to function if it doesn't get its supply of *each* of the nutrient types, so that "failure of a unit prevents nutrient flow through at least one of the channels serving that unit."

The difficulty should now be clear: we want to say that if the speaker has tacit knowledge of any theory then he must have tacit knowledge of T1, but the patterns of breakdown likely to occur in his linguistic behavior will suggest the ascription of tacit knowledge of T2. For example, if his competence with Gg is knocked out, because of the implications this has for the channels of nutrient flow, his competencies either with all other sentences containing G or with all other sentences containing g, or both, will be knocked out also.

		F	G	H	I	J	K	L	M	N	O
Xa	a	Fa	Ga	Ha	Ia	Ja	Ka	La	Ma	Na	Oa
Xb	b	Fb	Gb	Hb	lb	Jb	Kb	Lb	Mb	Nb	Ob
Xc	c	Fc	Gc	Hc	Ic	Jc	Kc	Lc	Mc	Nc	Oc
Xd	d	Fd	Gd	Hd	Id	Jd	Kd	Ld	Md	Nd	Od
Xe	e	Fe	Ge	He	Ie	je	Ke	Le	Me	Ne	Oe
Xf	f	Ff	Gf	Hf	If	Jf	Kf	Lf	Mf	Nf	Of
Xg	g	Fg	Gg	Hg	lg	Jg	Kg	Lg	Mg	Ng	Og
Xh	h	Fh	Gh	Hh	Ih	Jh	Kh	Lh	Mh	Nh	Oh
Xi	i	Fi	Gi	Hi	Ii	Ji	Ki	Li	Mi	Ni	Oi
Xj	j	Fj	Gj	Hj	Ij	Jj	Kj	Lj	Mj	Nj	Oj
		Y_F	Y_G	Y_H	Y_I	Y_J	Y_K	Y_L	Y_M	Y_N	Y_O

Figure 12.1

Davies's first point seems basically correct: not all causal structure will be germane to the attribution of tacit knowledge, so we need some account which will enable us to discriminate between causal structure that is thus relevant, and causal structure that is not. Nutritional structure is causal structure, but is intuitively irrelevant to the ascription of tacit knowledge. Why is this so? Davies suggests that this is because "the causal explanatory structure in the example is in no way *sensitive* to the *information* stored in the units," and that this lack of sensitivity is manifested in the fact that the patterns of *revision* of semantic beliefs are unlikely to follow the patterns of semantic decay: "we do not expect that *revision* of C's belief about the meaning of 'Fa' would go hand in hand with corresponding revisions of his beliefs about other sentences" (Davies, 1987, p. 453). So we need to revise our account of causal structure, and with it the mirror constraint, in such a way that the required sensitivity to informational content is introduced. In accordance with the remarks above, this is introduced via the notion of revision, so that the mirror constraint is modified thus:

> If, and only if, the operative states implicated in the causal explanation of a speaker's beliefs about the meanings of S_1, …, S_n are jointly sufficient for a causal explanation of his belief about the meaning of S; *and* those first states together with the revision of the speaker's belief about the meaning of S provide an explanation of the speaker's corresponding revisions in his beliefs about the meanings of S_1, …, S_n, should the semantic resources sufficient for the canonical derivation of truth-conditions specifications for S_1, …, S_n be sufficient for the canonical derivation of a truth-condition specification for S.

But does the introduction of the notion of sensitivity to information really solve the difficulty which Wright raised? I want to suggest that it does not, and that in fact Davies has misunderstood the character of Wright's objections here.

Wright's problem was *not* that some patterns of causal structure were irrelevant to tacit knowledge ascriptions, but that *even if* we could find a good constitutive account of tacit knowledge in which some form of causal structure was relevant to its ascription, this still would not justify articulation in the derivational structure of a semantic theory. Even if we *grant* the assumption that certain of the causal interrelations amongst speakers' competencies *are* worth describing, the objection raised is that we needn't run the risk of having the structure of these interrelations reflected indirectly via the derivational structure of a theory of meaning. Such structure could equally well be described (in the case of language L) by a listiform theory like T1, supplemented with a rider along the following lines: speakers are generally able to understand novel utterances provided they only involve familiar semantic primitives, and changes in their semantic beliefs about a sentence tend to be associated with changes in their semantic beliefs about all sentences containing one or more of the semantic primitives figuring in that sentence. More *detail* can then be obtained via the recursive syntax which was initially wedded to the listiform semantic theory. It might be worthwhile to pause briefly and investigate precisely how the relevant detail can be brought to light.

Wright suggests that the recursive *syntax* will provide this detail on the condition that it itself satisfies the mirror constraint. Now what does it mean to say that a syntax satisfies the mirror constraint? I suggest the following reconstrual of the mirror constraint for syntactical theories:

> If, and only if, the causal states implicated in the causal explanation of a speaker's beliefs about the meanings of S_1, …, S_n are jointly sufficient for a causal explanation of his belief about the meaning of S; *and* those first states together with the revision of the speaker's belief about the meaning of S provide an explanation of the speaker's corresponding

> revisions in his beliefs about the meanings of S_1, …, S_n, should the syntactic resources sufficient for the canonical derivation of well-formedness specifications for S_1, …, S_n be sufficient for the canonical derivation of a well-formedness specification for S.

If a syntactical theory satisfies this then *its* derivational structure (the canonical routes from its axioms to its specifications of well-formedness) will mirror with at least as much clarity the causal structure of speakers' competencies which was initially mirrored in the derivational structure of the semantic theory. It is important to note that this objection holds good even where the causal structure *is* sensitive to the informational content stored in the units of the representation of a semantic theory. Davies betrays his misunderstanding of this point in the following passage:

> It may be that by Wright's lights, no refinement of the notion of causal structure could ever justify the idea that a *semantic* theory should mirror that structure. But, to the extent that any refinement ensures that the salient causal structure can be described as an information-processing structure, this extreme view will be hard to sustain. (Davies, 1987, p. 454)

I disagree with this: even if we find that the salient causal structure can in fact be described as an information-processing structure, the objection still stands. The problem is not one of answering the question of how *causal* structure justifies articulation in a semantic theory, but rather of answering the question of how causal structure justifies articulation in a *semantic* theory.

Thus, notwithstanding the fact that Davies's revised account provides a useful sharpening-up of the notion of causal structure considered as relevant to the attribution of tacit knowledge, I would suggest that that revision leaves Wright's objection, properly read, completely untouched.

How, then, can we deal with Wright's objection? Wright's thought was that causal structure can be reflected by a purely syntactical – that is, non-semantic – theory, and that therefore some additional reasons have to be provided to ground the preference for reflection in theories of meaning. I think we can undercut Wright here by *denying* that he has shown how to reflect the salient causal structure in a non-semantic theory. Let's look more closely at the conjunction: listiform semantic theory plus syntactical theory which meets the mirror constraint. The latter part of the conjunction will reflect the same causal structure as any semantic theory which satisfies the mirror constraint. But is it really *non-semantic* in nature? I would say that it is not – that if we stipulate that the syntactical theory must meet the mirror constraint (as modified to apply to such theories) then it is no longer *purely* syntactical. Agreed, the theorems which form the output of this theory are concerned solely with the well-formedness or otherwise of the sentences of the language. But what a theory is *about*, what *sort* of theory a given theory *is*, is determined not only by the content of the sentences which make up its output, but also by the constraints in accordance with which that output is generated. Suppose, for example, we have a theory A whose output consists of sentences detailing the amount of money possessed by a sample of 100 Scottish women, and a theory B whose output consists of sentences detailing the amount of money possessed by a sample of 100 people, but which has been constructed in accordance with the constraint that the people included in the sample space should all be Scottish women. Then, even though theory B's output is couched in pronouncements of the form "Person X has £x" which feature no mention of Scotland or women, there is a clear sense in which that theory is still *about* Scottish women. To get back to the linguistic case, we should note that the mirror constraint is a *semantic* constraint – it makes reference to beliefs about the *meanings* of sentences and to relations that obtain between these beliefs. So, any theory

which is required to satisfy the mirror constraint will be a semantic theory to the extent that it will encode semantic information, including, despite appearances, any theory whose output mentions only the grammaticality of the language's sentences.

I thus deny that Wright has shown how the relevant causal structure can be reflected by a non-semantic theory: what he has given us is, in fact, an account of how that structure can be reflected in another *semantic* theory. Indeed, it is perhaps misleading even to speak of *another* semantic theory: just as the theory B above seems to be little more than a *reformulation* of the theory A, given the extensional equivalence of T1 and T2, the conjunction of T1 with a recursive "syntax" which is really partially semantic seems to me to be little more than a reformulation of the explicitly semantic T2.[12]

I now look at the third objection which Wright raises against Evans. According to Evans's original account, the job of the theory of meaning is to describe the dispositions which speakers have, corresponding to each of the expressions for which that theory provides a separate axiom. But if this is *all* that the theory is meant to do, then the twenty-first axiom of T2 – the compositional axiom – ought to be redundant, because someone who only has the dispositions described by the other 20 axioms will thereby be disposed to judge correctly of the truth-conditions of the sentences of L. This is again a consequence of the fact that the dispositions connected with the names and predicates are interdefined. However, without the compositional axiom the theory T2 will be paralyzed: it will be impossible to derive any truth-conditions specifications for the sentences of L. So "Evans has not shown how we are to construe an articulated semantic theory as a description of speakers' dispositions" (Davies, 1987, p. 444).

The way out of this difficulty is again to switch from talk of dispositions to talk of the states underlying those dispositions, and of the reflection of causal explanatory structure via the satisfaction of the mirror constraint. The job of the theory of meaning is now viewed not as the description of dispositions, but as the reflection of a certain sort of causal structure. There is, then, no obstacle preventing the theory T2 from satisfying the mirror constraint, and hence no obstacle to its reflecting the structure in question and thus doing its proper job.

Let me try to clarify this point with the aid of one of Davies's examples. Consider another semantic theory T3, which has the same axioms as T2 for the proper names of L, but differently styled axioms for its predicates. These axioms will instead be of the form:

> A sentence coupling a name with the predicate "F" is true iff the object denoted by that name is bald.

As Davies puts it, "What T3 does is to parcel out the content of T2's compositional axiom among the ten predicates of the language" (1987, p. 445), so that T3 is not open to Wright's objection even when we take the Evans line about the description of dispositions. More importantly, when we take the line in terms of causally operative states and reflection of causal structure, it seems that there is nothing to choose between the ascription of tacit knowledge of T2 and the ascription of tacit knowledge of T3. To see this, define a relation of *DS-equivalence* (equivalence in point of derivational structure) on theories of meaning as follows:

> Two (extensionally equivalent) theories Tk and Tm are DS-equivalent iff, for any sentences S_1, …, S_n, S: the semantic resources in Tk which suffice for the canonical derivation of truth-conditions specifications for S_1, …, S_n suffice also for the canonical derivation of a truth-condition specification for S if and only if the semantic resources in Tm which suffice

for the canonical derivation of truth-conditions specifications for S_1, …, S_n suffice also for the canonical derivation of a truth-condition specification for S.

Then it turns out that T2 and T3 are DS-equivalent, for "although T2 has an extra axiom relative to T3, the use of this resource in T2 is constant across all derivations of meaning-specifications for whole sentences" (Davies, 1987, pp. 446–447). I'll clarify this by means of an example. In T2 the semantic resources sufficient for the derivation of a truth-condition specification for "Ga" and "Fb" are sufficient also for the derivation of a truth-condition specification for "Fa":

Ga	(1)	"a" denotes Harry	(axiom for "a")
	(2)	an object satisfies "G" iff it is happy	(axiom for "G")
	(3)	a sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate	(compositional axiom)
	(4)	"Ga" is true iff Harry is happy	(from (1), (2), and (3))
Fb	(1)	"b" denotes John	(axiom for "b")
	(2)	an object satisfies "F" iff it is bald	(axiom for "F")
	(3)	a sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate	(compositional axiom)
	(4)	"Fb" is true iff John is bald	(from (1), (2), and (3))
Fa	(1)	"a" denotes Harry	(axiom for "a")
	(2)	an object satisfies "F" iff it is bald	(axiom for "F")
	(3)	a sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate	(compositional axiom)
	(4)	"Fa" is true iff Harry is bald	(from (1), (2), and (3))

We can see that the resources used in T2 in the derivations of the specifications for "Ga" and "Fb" were the axioms for "a," "b," "F," "G," and the compositional axiom. These give all we need in order to derive a specification for "Fa." Now, because of the way the compositional axiom is built into the axioms for the predicates in T3, the same thing holds in T3. Witness,

Ga	(1)	"a" denotes Harry	(axiom for "a")
	(2)	a sentence coupling a name with the predicate "G" is true iff the object denoted by the name is happy	(axiom for "G")
	(3)	"Ga" is true iff Harry is happy	(from (1) and (2))
Fb	(1)	"b" denotes John	(axiom for "b")
	(2)	a sentence coupling a name with the predicate "F" is true iff the object denoted by the name is bald	(axiom for "F")
	(3)	"Fb" is true iff John is bald	(from (1) and (2))
Fa	(1)	"a" denotes Harry	(axiom for "b")
	(2)	a sentence coupling a name with the predicate "F" is true iff the object denoted by the name is bald	(axiom for "F")
	(3)	"Fa" is true iff Harry is bald	(from (1) and (2))

Here the resources used in the derivations of specifications for "Ga" and "Fb" were the axioms for "a," "b," "F," and "G." And again, these give us all we need in order to derive a specification for "Fa." We could do this again for all the appropriate sentences of L, and this would amount to a conclusive proof of Davies's assertion that T2 and T3 are DS-equivalent.

We thus find ourselves in the following position: because of the DS-equivalence of T2 and T3, either both theories satisfy the mirror constraint for a given speaker, or neither does; so that the 5% difference – the difference between the 20 axioms of T3 and the 21 axioms of T2 – *doesn't* in fact matter. Given that we are not concerned with the description of dispositional states apart from their position in a causal web whose structure we are concerned to reflect, the fact that we can describe that causal structure in causal theories which do not devote an individual axiom to name–predicate concatenation is innocuous: it does not vitiate its reflection in theories which *do* contain such an axiom.

However, as Davies realizes, there is now another objection in the offing. Suppose we are trying to decide whether a given articulated semantic theory satisfies the mirror constraint with respect to a particular speaker of L. Suppose also that that speaker revises his belief concerning "Fb" from "John is bald" to "John is baldish." In accordance with the mirror constraint, we will check whether he revises his beliefs about "Fa," "Ga," and "Gb" *correspondingly*. But what counts as the *corresponding* revision of, say, "Gb"? Is it the null revision, which leaves the speaker with the belief that "Gb" means that John is happy, or is it the revision which leaves him with the belief that John is happyish? No one of these answers seems to be uniquely correct, and according to Davies this latent indeterminacy is a time-bomb which threatens the stability of his proposed account of tacit knowledge. But this threat seems to dissipate somewhat when we note that the indeterminacy can in fact be resolved "according as we look at the semantic properties of a language through the grid of one theory rather than the other" (Davies, 1987, p. 459). Precisely how this resolution is achieved can be seen from the following quotation from an earlier paper of Davies's – to say that the revision of the semantic belief concerned with S_i corresponds (considered from the viewpoint of a particular semantic theory) to the revision in the semantic belief concerned with S is to say that:

> If A were to revise his belief about the meaning of S in that respect of the meaning which the semantic theorist discerns as a deductive consequence [in the semantic theory in question] of the presence in S of the syntactic item Γ assigned the semantic property Δ, and if what A believed about the meaning of S as the result of this revision were to be the deductive consequence of a revision of Δ to Δ^*, then A would revise his belief about the meaning of S_i (and the meanings of any other sentences containing Γ) in such a way as would be the deductive consequence (in the theory in question) of the assignment to Γ of Δ^* rather than Δ. (Davies, 1981b, p. 149)

Given this, we can now say which revisions merit the ascription of tacit knowledge of T2, and which merit the ascription of T3. And it seems that these will not coincide:

> If we consider T2, then there are two possible changes in the proper axioms, each of which would result in "Fb" being assigned the meaning that John is baldish. A change in T2's axiom for "F" has one pattern of consequences: a change in the 21st axiom – the compositional

> axiom – has a different, and more extensive pattern of consequences. If we consider T3, on the other hand, then there is only one possible change to the proper axioms which would result in "Fb" being assigned the meaning that John is baldish. (Davies, 1987, p. 459)

So, our original account, in which the ascription of tacit knowledge of T2 simply was ascription of tacit knowledge of T3, and vice versa, will have to be revised in such a way as to take this into account. For consider the following two speakers A and B

> [f]or whom, a form of the language of thought hypothesis is true. For these speakers, language comprehension – in particular, the assignment of meaning to sentences – is a matter of derivations in a semantic theory explicitly represented in a special purpose language processing system. Suppose that speaker A conducts on his inner blackboard derivations in theory T2, while speaker B conducts derivations in theory T3 … For the purposes of tacit knowledge ascriptions speakers A and B are grouped together. (Davies, 1987, pp. 447–448)

But now:

> Just as theory T2 with its 21 axioms provides an extra locus of content sensitivity over theory T3 with its 20 axioms, so the causal explanatory structure in speaker A provides an extra locus of systematic revision over the causal explanatory structure in speaker B. So, not altogether surprisingly, it is speaker A – conducting inner derivations in theory T2 – who meets the condition for tacit knowledge if the indeterminacy is resolved by looking at the language through the grid of T2. And it is speaker B who meets the condition if the indeterminacy is resolved by looking at the language through the grid of theory T3. (Davies, 1987, p. 459)

What is going on here? The suggestion is that *only if* we have a one-to-one correspondence between the axioms of a theory of meaning and the explanatory loci of systematic revision which go towards making up the causal explanatory structure in a speaker can we spell out satisfactorily the notion of systematic revision amongst the implicit semantic beliefs that constitute that speaker's linguistic competence. For if the axioms for the language's expressions and the speaker's competencies with those expressions correspond with each other one by one, then since "for each axiom or rule, the required notion of *systematic revision* can be spelled out in a quite determinate way," similarly, determinate sense can be made of the notion of systematic changes in the nature of the speaker's competencies, that is, in his revisions of his semantic beliefs. If this is correct then the 5% difference *will* matter – A will be viewed as having tacit knowledge of T2 (and not of T3) while B will be viewed as having tacit knowledge of T3 (and not of T2).

Although I am suspicious of the details of Davies's example of the speakers A and B – it is unclear, for instance, whether A really has the 21 causal explanatory states required for tacit knowledge of T2 (what state now corresponds to the compositional axiom?), and witness the hardly uncontroversial assumption about the language-of-thought hypothesis – I think we can accept that he has provided a good argument to the effect that a speaker with causal states underlying Evans's dispositions can only be ascribed tacit knowledge of T3 and not T2, because in the latter case we would have no means of resolving the indeterminacy which surrounds the possible revisions which such a speaker could make in his semantic beliefs. However, as Davies himself points out, this amounts only to a minor revision, and certainly not to a wholesale rejection, of the account of tacit knowledge which has its roots in Evans's suggestions. As Davies puts it, "the form of Evans's original proposal shines through" (1987, p. 461).

5 The Mirror Constraint and Understanding Novel Utterances

Can the imposition of the mirror constraint help us understand speakers' capacities to understand novel utterances, provided they include only familiar semantic primitives and modes of construction? Suppose that a speaker can come to understand S after exposure to the sentences S_1, …, S_n, that loss of competence with S occasions loss of competence with at least some of S_1, …, S_n, and that a revision of a speaker's belief concerning the meaning of S occasions corresponding revisions in his beliefs about the meanings of each of S_1, …, S_n. Then, on the basis of this evidence we shall claim that: the operative states implicated in the causal explanation of the speaker's beliefs about S_1, …, S_n are jointly sufficient for a causal explanation of his belief about the meaning of S, and those first states, together with his revision of his belief about the meaning of S, are sufficient for an explanation of the corresponding revisions in his beliefs about the meanings of S_1, …, S_n. Of course, to make such a claim is not yet to *provide* the causal explanation alluded to, so that we are still left with the question: How is it possible that the operative states implicated could be thus sufficient? It is precisely this question that is answered by showing how the semantic resources sufficient for the canonical derivation of meaning specifications for S_1, …, S_n are sufficient for the canonical derivation of a meaning specification for S, by the provision of a semantic theory in which they are so sufficient. That is, the move from the operative states to the belief about the content of the novel utterance is viewed as being explained by the provision of a theory which generates, on the basis of the information represented by the relevant operative states – represented by the axioms of the theory – a meaning-specifying theorem which gives the content of the novel utterance in question. So, the provision of a compositional theory of meaning is supposed to make explicable the notion that a speaker could come to form an implicit belief about the content of a novel utterance on the basis of a pre-existing set of causally operative states, because the route from the operative states to the belief about content is reflected by the derivational route from the axiomatic base of the semantic theory to the appropriate meaning-specification. Thus, the construction of a compositional semantic theory, and the notions that speakers might have, in the requisite sense, tacit knowledge of such a theory, is supposed to help answer the question as to how it is possible for speakers to understand utterances of previously unencountered sentences.[13] Davies thus goes beyond Evans, who claims that "the notion of tacit knowledge of a structure reflecting theory of meaning, explained as I have explained it, cannot be used to explain the capacity to understand new sentences" (1981a, p. 134). I shall return to this difference between Evans and Davies in the final section.

6 Wright's Proposal

In addition to the objections responded to in §4, Wright complains that whereas the account of tacit knowledge given by Evans – and by implication, the development of that account via the imposition of Davies's mirror constraint – requires the semantic theorist to pay attention to empirical facts about language acquisition, loss, and revision, *actual* semantic theorizing seems to have proceeded in happy ignorance of such facts. Loath to conclude that the right conception of the semantic theorist's task "has not greatly impinged upon the consciousness of workers in the field," Wright proposes an alternative account of what theorists of meaning ought to be doing, which is claimed to harmonize better with their actual *modus operandi*. In this section I briefly outline Wright's alternative proposal, and argue that it

presupposes, rather than undercuts, the Evans–Davies account which proceeds via the imposition of the mirror constraint.

Speakers are able to understand novel utterances provided they include only familiar semantic primitives; and the semantic theorist's task is to give an answer to the question, "How do they do it?" Wright sees this question as dividing into two sub-questions: (a) How is it possible for there to be a device which could, when fed with information concerning the visible or audible structure of a sentence, process that information in such a way that a theorem about the meaning of the sentence is generated? and (b) How is it possible for such a device actually to be embodied in normal, competent speakers of a natural language? Question (a) is conceived of as being answerable *a priori*, independently of the empirical evidence upon which Evans and Davies lay so much stress: it will have been answered when a suitable "computer program" has been written, the writing of which will not demand elevation from the armchair. Wright thus thinks that if we view semantic theorists as attempting to answer question (a) in the manner he describes, the discrepancy between the account of what their project is, and what they actually do, will vanish. And he views the provision of an answer to (a) as an essential prerequisite for any attempted answer to (b):

> The sort of understanding of the actual capacities of speakers which is called for would be achieved exactly when enough was known about them to enable us to understand how in detail they embody such a device. And, of course, there can be no such understanding before we have formed the appropriate theoretical conception of the powers which the device must have. Doing that requires writing the computer program. (Wright, 1986, p. 236)

However, it seems to me that Wright has overlooked something crucial here. Given that we are interested in answering the question "How do *they* do it?" even after (a) and (b) have been answered there remains the further question: (c) Do they, in reality, embody the device whose theoretical powers the computer program describes? An affirmative answer to this question is crucial if the answers given to (a) and (b) are to have any explanatory value: an account of the theoretical powers of the device and an account of how it might be *possible* for speakers to embody it will have no explanatory value as an answer to the question "How do they do it?" if they do not actually embody the device in question. Questions (a) and (b) together tell us how it *could* be done, but what we are really after is an account of how it *is* done, and for this we need an answer to (c). And, of course, the provision of an answer to (c) is an empirical matter, dependent on the sort of empirical evidence concerning language acquisition and loss Evans and Davies focus on: whether or not a speaker actually does embody a particular information-processing mechanism is a matter which is amenable to empirical investigation. We will not view a particular speaker as embodying a particular information-processing mechanism unless the derivational structure of that mechanism is isomorphic to the causal explanatory structure, the existence of which is suggested by the appropriate sorts of empirical data.

So if semantic theorists are out to answer the question "How do they do it?" of our capacity to understand novel utterances, they are at some point going to have to pay attention to empirical detail. The question is, of course, where? Two broad answers to the question suggest themselves. The semantic theorist could begin to pay attention to such detail at the final stage of his enterprise: having written the appropriate computer program, and having shown – somehow or other – that it is possible for it to be embodied in speakers, he could then *go on* to ask whether or not they actually do embody it. And it is at this final stage that the usual sources of empirical evidence will be crucial. The semantic theorist would thus be

attempting to answer the questions (a), (b), and (c) in a straightforwardly linear fashion. But it is clear that this could be an extremely inefficient way of going about things: once answers to (a) and (b) have been provided, there is no guarantee that the required affirmative answer to (c) will be forthcoming. The semantic theorist could write out his program, argue that it is possible for the information-processing device to be embodied in actual speakers, and then discover – to his horror – that it is in fact *not* actually embodied. And this could happen time and time and time again. Of course, he might get an affirmative answer to (c) on one of his early attempts after all; but it is clear that if he is proceeding in the linear fashion described, this could only have the status of a happy accident.

An alternative, and much more efficient, way of proceeding would be to write the computer program out as before, but this time in accordance with empirical constraints which ensure that, once (a) and (b) have been answered, no obstacle remains to the production of an affirmative answer to (c). In effect, this involves writing the computer program – constructing the theory of meaning – in accordance with some constraint along the lines of the mirror constraint. The semantic theorist is thus still viewed as attempting to answer the questions (a), (b), and (c), but no longer in the simple linear fashion described above. We answer (a) and (b) subject to the constraint that there is an affirmative answer to (c) via the imposition of something like the mirror constraint, and we are justified in so doing because without an affirmative answer to (c) the answers to (a) and (b) will have no explanatory value as answers to the question, "How do they do it?" There is no point in writing a program describing the theoretical powers of an information-processing device which we do not embody if the whole aim of the enterprise is to achieve an understanding of how *we* can understand novel utterances.[14]

My conclusion is thus that, far from providing an account of the semantic theorist's task which does not involve the imposition of the mirror constraint, Wright's suggestions, in the end, only serve to highlight the need for the imposition of that very constraint.[15]

7 Tacit Knowledge and Rule-Following

In the above, I have been attempting to defend the suggestion that the construction of theories of meaning should be subject to Davies's mirror constraint. But there is at least one major problem outstanding which anyone wishing to embrace the account I have defended would have to face up to. I cannot do more than briefly mention this problem here. It stems from the fact that our account attempts to give an explanation of language mastery – in particular of the capacity to understand novel utterances – in *cognitive-psychological* terms. In Wright's words, it is an ability we are conceived to have

> because we are appropriately related to a finite body of *information* which may be inferentially manipulated in such a way as to entail, for each novel string on which we can exercise our "linguistic-creative" power, appropriate theorems concerning its grammaticalness and content,

which commits us to

> the picture of language as a kind of syntactico-semantic mechanism, our largely unconscious knowledge of which enables us to compute the content which, independently and in advance

> of any response of ours, it bestows on each ingredient sentence … [and in which] the mechanism does the generating and the competent adult keeps track of what (and how) it generates. (Wright, 1989, pp. 233 and 238)

Prima facie, this seems like an accurate description of our account: the information codified in the causal states corresponding to semantic axioms is conceived of as settling in advance, and independently of anything we go on to say or do, the content of the totality of admissible sentences in the language.

But why should this be a problem? What is wrong with the picture of language as a syntactico-semantic mechanism whose output competent speakers are able to track? Wright's suggestion is that it is precisely this sort of picture of the ability constitutive of language mastery which is the ultimate target of Wittgenstein's remarks in the *Investigations* and elsewhere on the nature of rule-following: that if we try to construe language mastery as an ability to track states of affairs constituted independently and in advance of what we say or do, we shall find ourselves unable to give any coherent account of the epistemology of the tracking accomplishment. (For further discussion, see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §4.) Thus, the question here is: *Does the account of tacit knowledge which uses the mirror constraint actually commit us, as Wright suggests, to conceptions of meaning and linguistic understanding which the rule-following considerations would counsel us against?* Perhaps one way to try to avoid an affirmative answer to this question would be to retain Davies's version of the mirror constraint as it stands, but drop the claim that the causal explanatory states corresponding to semantic axioms have informational content – that they represent information – which somehow settles in advance the content of the intentional states constitutive of sentential understanding. It is difficult to assess this suggestion in the absence of an account of what the relationship between the causal states and the axioms is, and of when causal states which are not genuinely intentional can be claimed to have informational content.[16] John Campbell takes Evans to be suggesting that the causal states underlying language mastery have no such content:

> [Evans suggests that] the discernment of structure by a description of understanding is in effect a discernment of non-psychic, purely neural structure … [and that] there is no *psychological* machinery which typically explains a speaker's perception of the meaning of the heard sentence. (Campbell, 1982, p. 24)

Davies, on the other hand, is explicit in his desire to take the opposing view:

> We ought to explore the differences between propositional attitudes *and other information-containing cognitive states*. (Davies, 1986, p. 140; see also his 1989)

Deciding who is right would require a full account of the sort of content possessed by "subdoxastic states," together with an account of the conditions under which such content should be ascribed. This question should, perhaps, be the starting point for any future discussion of the notion of tacit knowledge. Rather than address that here, though, I'll finish by pointing the reader in the direction of a series of papers in which I argue that Wright's own account of the upshot of the rule-following considerations is consistent with the idea that the underlying causal states mentioned in the mirror constraint can be viewed as possessing informational content in a way that would subserve a cognitive-psychological explanation of our capacity to understand novel utterances.[17]

Notes

1 We should distinguish between two senses of the phrase "theory of meaning." This could signify, on the one hand, a theory relating to a single language which attempts to state the meaning of every sentence in that language; or, on the other, a theory relating to language in general, which "attempts to analyse, elucidate, or determine the empirical content of the concept of meaning in general" (Sainsbury, 1979, p. 127). In this chapter we will be concerned exclusively with theories of meaning in the former sense. I shall assume that a natural language is one which is learned by training and projection from that training.

2 In what follows I assume that the meaning-specifications have the form of truth-condition specifications, in the manner of essays 1–5 of Davidson (1984). (See also Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S.) This assumption has been widely disputed: see Dummett (1975; 1976). But I do not enter into these issues in this chapter: for our present purposes, nothing of importance hinges on the outcome of that particular debate.

3 Stephen Stich (1978) suggests a necessary condition on a cognitive state's counting as a genuine belief which is very similar in spirit to the constraint which we just extracted from the Evans–Wright discussion. Stich uses the notions of *inferential integration* and *insulation*. Whereas genuine beliefs form inferentially integrated subsystems of an agent's cognitive states (in the sense that there are *many* inferential routes, both deductive and inductive, from a given belief to any other belief), *subdoxastic* states, in contrast, are inferentially insulated from the vast majority of the subjects' genuine beliefs, in the sense that a subdoxastic state will be linked inferentially to only a very limited range of the agent's beliefs.

4 Might there be other reasons for finding the notion of tacit knowledge of meaning-delivering theorems problematic? How could the theorems encapsulate a rule governing the use of sentences, followed even by competent speakers who do not register the theorem in consciousness? As Wright puts it (1986, p. 218), "How can a principle function as a rule if those who engage in the practice which it is supposed to regulate have no consciousness of it?" Wright convincingly rejects this line of objection to tacit knowledge of theorems. In addition, we might add that if Wittgenstein's rule-following considerations have taught us anything, it is that the following question is no easier to answer: How can a principle function as a rule if those who engage in the practice which it is supposed to regulate *are* in fact conscious of it?

5 Note that it would be a mistake to think that a similar argument could reinstate tacit knowledge of a meaning-theoretic axiom as a genuine propositional attitude, since tacit knowledge of an axiom A can equally interact with a multiplicity of other states, namely, those corresponding to the sentences in which the expression governed by A appears. There is no real analogy with the example concerning genuine beliefs, since the states corresponding to the axioms will be implicated only in the causal explanation of the agent's possession of the states corresponding to the theorems. No such implication need hold in the belief example: for example, the belief that P might not play any part in the causal generation of the belief that $P \rightarrow Q$. ("Interact," as it appears in Constraint 2, really concerns *rationalistic* explanation: to say that state A interacts with state B to lead to state C is to say that state A and state B together rationalistically explain the presence of the state C. The crucial point about the axiom-theorem case will then be that the citation of the state corresponding to the axiom will be redundant, so far as rationalistic explanation is concerned (as I go on to argue in the text).)

6 Hess took two photographs of the same girl, enlarged the pupils of her eyes in only one of them, and then showed the two pictures to a range of male subjects who were unaware of the change. Hess found that the males consistently described the girl as looking more attractive in the altered picture, although they were unable to say in what the difference consisted. The idea is that there must have been causal states of the males containing information about relative pupil size which played a part in the causal production of the explicit beliefs about relative attractiveness, although those causal states are not to be counted as genuine beliefs in themselves. For a fuller discussion of the Hess experiment, see Stich (1978, pp. 503, 505–506, and 511).

7 In correspondence, Lars Dänzer has suggested that although tacit knowledge of a semantic axiom never figures non-redundantly in a rationalistic explanation of linguistic action, nevertheless, since beliefs about word meaning *can* figure non-redundantly in rationalistic explanations of beliefs about sentence meaning, there are still grounds for viewing knowledge of a semantic axiom as a genuine intentional state. This is an interesting suggestion, but I'm unconvinced for the following reason. As we'll see more clearly below, tacit knowledge of an axiom such as

(1) "a" denotes Harry

is to be viewed as underlying competent use of sentences in which the name "a" is *used* (such as e.g., the sentence "a is F"). However, a belief about the meaning of "a," on Dänzer's suggestion, is implicated in the rationalistic explanation of a speaker's acceptance of a sentence in which the name "a" is *mentioned* rather than used – for example, his acceptance of the sentence "'a is F' means that Harry is bald." So the fact that a belief about word meaning can figure in this latter sort of explanation does not license the construal of tacit knowledge of a meaning-theoretic axiom as a genuine propositional attitude. The distinction on which this point turns is perhaps masked by the fact that the formulations of the mirror constraint given below are couched in terms of a speaker's beliefs *about* the meanings of sentences. In these formulations, though, "a speaker's belief about the meaning of a sentence S" really refers to the cognitive-psychological state implicated in his understanding *of* S rather than to a belief that potentially rationalizes his acceptance of sentences in which S is mentioned. (For reasons similar to these, I suspect that Bernhard Weiss's (2004) argument that – *contra* Evans and Wright – knowledge of word meaning can count as a genuine propositional attitude fails to establish that tacit knowledge of a semantic axiom can also be viewed as a genuine propositional attitude.)

8 Note that the analogue of the listiform theory for a language with an infinite number of sentences is provided by an infinitary axiom schema: A is T iff P, "where 'P' may be replaced by any declarative sentence of the object language and "A" by the quotational name of that sentence" (Wright, 1986, p. 211).

9 There's an obvious connection here with Gareth Evans's "Generality Constraint": see Evans (1982, pp. 100–105).

10 We ought to note, in fairness to Evans, that this seems to be precisely what he was after in the first place – witness his claim that dispositions have to be given a full-blooded characterization.

11 Note that this holism does not imply that we shall not be able to pair explanatory states with individual axioms.

12 Darragh Byrne (2004, p. 73) attempts to criticize the argument of mine outlined in this and the previous paragraph. He suggests that Wright's objection is that Evans has not done enough to justify the idea that the states corresponding to the axioms of T2 have genuine representational contents. He argues that it is irrelevant that the recursive syntax that satisfies the mirror constraint is "semantic in nature": "[T]he 'supplement' which Wright suggests might explain the structure underpinning the speaker's competence is not this mirror-constraint-satisfying syntactic theory *itself* – the 'supplement' is *based* on that syntactic theory" (2004, p. 73, Byrne's emphases). It's not obvious to me that Byrne's characterization of Wright's objection is correct, but even if we assume that it is, his remarks are unconvincing. Wright clearly takes the syntactic theory itself to provide the necessary detail about the relevant causal structure. He writes: "Admittedly, such a rider [what Byrne calls the 'supplement' – AM] would not be a detailed, or axiomatic, description of the interrelations which the Mirror Constraint would have a theory of meaning reflect. *But there is every reason to think that the recursive syntax which the theorist adjoins to his infinitary semantic theory would supply the materials for the more specific descriptive task*" (Wright, 1986, pp. 213–214, my italics in the second sentence). Wright clearly thinks that the detail on the causal structure is provided by the *syntax* rather than the rider ("supplement"). And even if Byrne were correct in thinking that it was the rider that provided the relevant detail, his admission that it is *based* on an explicitly syntactical (but implicitly semantical) theory would leave my argument against Wright intact.

13 Given my remarks at the start of §2, this shows that the construction of semantic theories in accordance with the mirror constraint can also help to answer the question about learnability.
14 Louise Antony independently runs a similar line of argument in her (1997).
15 So although I do not want to go so far as to claim that "the right account has not greatly impinged upon the consciousness of workers in the field," I would suggest that the most efficient way of implementing that account seems to have eluded many of its practitioners.
16 Note, though, that if we take this line we appear to lose the possibility of explaining, in cognitive-psychological terms, our capacity to understand novel utterances.
17 The relevant papers are Miller (2007; 2012; 2013). A fuller discussion of tacit knowledge would also have to include a discussion of "the doctrine of essential linguistic structure," as discussed in Fricker (1981) and Sainsbury (1979). I discuss this issue – doubtless not at the length it deserves – in Miller (2012, §4). I would like to thank Michael Clark, Lars Dänzer, Martin Davies, John Divers, Jim Edwards, Bob Hale, Bob Kirk, Greg McCulloch, Joe Mendola, Stephen Read, Ali Saboohi, Mark Sainsbury, Laura Schroeter, Roger Squires, Jim Stuart, and Crispin Wright for very helpful comments and discussion.

References

Antony, L. 1997. "Meaning and semantic knowledge." *Proceedings of the Aristotelian Society*, suppl. vol. 71: 177–209.

Byrne, D. 2004. "Three conceptions of tacit knowledge." *Agora: Papeles de Filosofia*, 23: 61–85.

Campbell, J. 1982. "Knowledge and understanding." *Philosophical Quarterly*, 32: 17–34.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davies, M. 1981a. *Meaning, Quantification, and Necessity: Themes in Philosophical Logic*. London: Routledge.

Davies, M. 1981b. "Meaning, structure, and understanding," *Synthese* 48(1): 135–161.

Davies, M. 1986. "Tacit knowledge and the structure of thought and language." In *Meaning and Interpretation*, edited by C. Travis, pp. 127–158. Oxford: Blackwell.

Davies, M. 1987. "Tacit knowledge and semantic theory: can a 5% difference matter?" *Mind*, 96(384): 441–462.

Davies, M. 1989. "Tacit knowledge and subdoxastic states." In *Reflections on Chomsky*, edited by A. George, pp. 131–152. Oxford: Blackwell.

Dummett, M. 1975. "What is a theory of meaning?" In *Mind & Language*, edited by S. Guttenplan, pp. 97–138. Oxford: Oxford University Press.

Dummett, M. 1976. "What is a theory of meaning? (2)" In *Truth and Meaning*, edited by G. Evans and J. McDowell, pp. 67–137. Oxford: Oxford University Press.

Evans, G. 1981. "Semantic theory and tacit knowledge." In *Wittgenstein: To Follow a Rule*, edited by S. Holtzmann and C. Leich, pp. 118–137. London: Routledge.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fricker, E. 1981. "Semantic structure and speakers' understanding." *Proceedings of the Aristotelian Society*, 83: 49–66.

Miller, A. 2007. "Rules-as-rails, tacit knowledge and semantic creativity." *International Journal of Philosophical Studies*, 15(1): 125–140.

Miller, A. 2012. "Judgement dependence, tacit knowledge and linguistic understanding." In Stalmaszczyk, 2012, pp. 405–428.

Miller, A. 2013. "The development of theories of meaning." In *A Handbook of the History of Analytic Philosophy*, edited by M. Beaney, pp. 656–688. Oxford: Oxford University Press.

Sainsbury, R. M. 1979. "Understanding and theories of meaning." *Proceedings of the Aristotelian Society*, 80: 127–144.

Stalmaszczyk, P., ed. 2012. *Philosophical and Formal Approaches to Linguistic Analysis*. Berlin: Ontos Verlag/Walter de Gruyter.

Stich, S. 1978. "Beliefs and subdoxastic states." *Philosophy of Science*, 45(4): 499–518.

Weiss, B. 2004. "Knowledge of meaning." *Proceedings of the Aristotelian Society*, 104: 75–94.

Wright, C. 1980. *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.

Wright, C. 1981. "Rule-following, objectivity, and the theory of meaning." In *Wittgenstein: To Follow a Rule*, edited by S. Holtzmann and C. Leich, pp. 99–117. London: Routledge.

Wright, C. 1986. "Theories of meaning and speakers' knowledge." In *Realism, Meaning, and Truth*, pp. 204–238. Oxford: Blackwell.

Wright, C. 1988. "How can the theory of meaning be a philosophical project?" *Mind & Language*, 1(1): 31–44.

Wright, C. 1989. "Wittgenstein's rule-following considerations and the central project of theoretical linguistics." In *Reflections on Chomsky*, edited by A. George, pp. 233–264. Chichester: Wiley-Blackwell.

Further Reading

Devitt, M. 2006. *Ignorance of Language*. Oxford: Oxford University Press.

Gascoigne, N., and T. Thornton. 2013. *Tacit Knowledge*. Durham, NC: Acumen.

Gillett, G. 1988. "Tacit semantics." *Philosophical Investigations*, 11(1): 1–12.

Knowles, J. 2000. "Knowledge of grammar as a propositional attitude." *Philosophical Psychology*, 13(3): 325–353.

Miller, A. 2014. "Review of Gascoigne and Thornton *Tacit Knowledge*." *International Journal of Philosophical Studies*, 22(4): 630–635.

Rattan, G. 2002. "Tacit knowledge of grammar: a reply to Knowles." *Philosophical Psychology*, 15(2): 135–154.

Saboohi, A. 2012. "In defense of implicit knowledge in a full-blooded theory of meaning." In Stalmaszczyk, 2012.

13

Radical Interpretation

JANE HEAL

1 A Bird's-Eye View of Some Options

To engage in radical interpretation is to set about investigating the meanings of utterances in some completely unknown language. It has been suggested that reflection on how such interpretation should proceed will throw light on the nature of meaning. The most influential proponent of this idea is Donald Davidson (1984, especially essays 9–12). This chapter will therefore be much concerned with his proposals. But it aims also to locate his views in a broader context and to consider alternative approaches.

The structure of this chapter is as follows. The remainder of this section discusses the location of radical interpretation within the broader field of philosophy, and identifies some of the options and their presuppositions. §2 outlines the ideas of Davidson; and §§3, 4, and 5 consider their contrasts with alternative views, seeking to identify the crucial issues.

Two things need clarification at the start. The first is that the epistemological appearance of the enterprise, if taken too seriously, could be misleading. To think that we should get illumination on the ontology or metaphysics of some concept by asking how judgments using it are established is characteristically an empiricist view. But talk of "what one would need to know in order to establish such-and-such" may also be a mere rhetorical device for the vivid presentation of independently motivated metaphysical proposals. We should regard our reflections on the imagined procedures of radical interpretation in the second way and not the first. This is clearly the approach of Davidson and others; the investigations they speak of are highly idealized, and neither they nor we wish to commit ourselves at the outset to controversial aspects of empiricism (see Davidson, 1990, and Lewis, 1983, pp. 110–111).

A second point needing early clarification is that our quarry is not just the notion of linguistic meaning but also a broader notion of meaning or representational content, in which such content may be attributed to psychological states as well as to linguistic items. We are to think about what fixes the content of a person's thoughts as well as about how we

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

could identify the meaning of what he or she says. Davidson takes it that thought cannot occur without language, and that language is the primary vehicle of thought (1984, essay 11). So for him it follows immediately that investigation of the nature of thought and investigation of the nature of language are one and the same. But even those philosophers who do not accept this are likely to agree that thought and language are closely linked. It seems impossible that, for a language-using creature, there could be two entirely independent sets of facts, one about what he or she thinks and one about the meaning of what he or she says.

What exactly is 'radical interpretation'? And what is presupposed by the idea that it is possible? Its proponents seem to mean by the phrase an inferential process which starts from information, *all* of which is non-semantic, and ends with attribution of rich and varied meanings. Thus Davidson says that radical interpretation must start from "evidence that can be stated without essential use of such linguistic concepts as meaning, interpretation, synonymy and the like" (1984, p. 128). Lewis characterizes the matter thus:

> At the outset we know nothing about [our subject's] beliefs, desires and meanings … Our knowledge … is limited to our knowledge of him as a physical system … Now, how can we get from that knowledge to the knowledge we want [sc. the knowledge of meanings]? (Lewis, 1983, p. 108)

To give this description is not to say that radical interpretation cannot have among its starting points some general principles about meaning, in the form of explicit or implicit instructions on how to process the non-semantic information presented. Clearly without such general principles we should be completely hamstrung. But the starting information is not to contain any attributions of actual particular meanings, even to expressions of one's own language, since on Davidson's fully developed theory our knowledge of the meanings in our own first language is based on radical interpretation.

We can see why this restriction on the nature of the evidence might be imposed by someone who wishes to use reflection on radical interpretation to cast light on the metaphysics of meaning, that is, on the nature of meaning as a phenomenon and on its relation to other kinds of fact. If we imagine ourselves interpreting non-radically – inferring to meanings from a mixed body of knowledge containing facts about meaning as well as non-semantic facts – it may be that we can, on reflection, discover some distinctive patterns of relation between our non-semantic premises and our conclusions. It will, however, be difficult to build much upon this for metaphysical purposes (to use it, for example, to build theories of the relation of the semantic to the physical) until we are clear what has been the distinctive contribution of the additional semantic premises. But it is precisely the nature of the semantic which we are trying to clarify. And thus we risk going round in a circle. The claim here is not that any philosophical enterprise of this shape is bound to be hopeless; the claim is only that it is apparent that metaphysical conclusions could much more easily be drawn if the inferences were 'radical' in the sense outlined.

We can now see clearly at least one presupposition of the idea that radical interpretation is possible; it is that rich meaning notions are not an essential part of the basic observational vocabulary with which we approach the world. We may speak colloquially of hearing a person say that something is the case, or seeing that a person is thinking such-and-such. But, says the believer in radical interpretation, this is a mere useful idiom, not to be taken seriously. What is really observationally apparent must be something less committal, for example that certain sounds or movements were produced in certain patterns, having such-and-such causes, and the like. Facts about meaning are somehow based on or inferred from such facts.

Let us for the moment accept that there could be such a radically non-semantic starting point as the one imagined. There are now broadly two possibilities for how the attempted working-out of meanings might go. First we could maintain a form of dualism. On this view, possession of meaning by a physical vehicle is a matter of that vehicle being suitably related to some unobservable entity, or itself having some unobservable property. Perhaps, for example, it is the effect of some intrinsically representational state of a Cartesian mind. (This is the theory of substance dualism.) Or perhaps it has in itself another hidden or mental aspect. (This would be the theory of a 'double-aspect' property dualism.) The investigation of meanings on this hypothesis is very similar to the investigation of micro-organisms or invisible dwarf companions to visible stars. Only it is not quite like this, because dualistically conceived meanings are also thought of as *in principle* unobservable, even with microscopes or spaceships. They are thought of as causing public manifestations, but being essentially linked, in a distinctive and epistemologically privileged way, to their subject. There are many familiar lines of argument against dualist positions, for example the fact that they produce 'other minds' problems and that they make it difficult to give an intelligible account of mind–body interactions.

Those philosophers who write on radical interpretation and whose work we shall consider take it for granted, however, that dualism is not a serious option. This means that, for them, the other possibility is the one which must be accepted. This possibility is that all the facts that are relevant to meaning are there in the non-semantic (that is, the physical or material) assemblage. Meaning is not, as in dualism, something independent of this assemblage but is, on the contrary, something fixed and constituted (in so far as it is fixed and constituted) by the non-semantic. So any materialist theory of mind which is non-eliminativist and which addresses itself seriously to the question of intentional content is a theory of radical interpretation.

It is not the case that the believer in the possibility of radical interpretation who also rejects dualism is committed to any simple-minded form of reductionism. He or she is not committed, for example, to the idea that semantic statements can be translated or unpacked one by one into packages of non-semantic ones. As we shall see, Davidson's proposal is very different from anything of this kind. But it is the case that the believer in radical interpretation is committed to the idea that the semantic arises from, or is constituted by, some kind of appropriate complexity in the non-semantic. For want of a better word, I shall say that he or she is committed to the reducibility of the semantic to the non-semantic. But it is to be remembered that what is involved is reducibility in some extremely broad sense.

The difficulties of dualism have given a bad name to the whole idea of non-reductive accounts of meaning (in the very broad sense of 'reduction' just gestured at). The bulk of philosophical writing on meaning (in the analytic tradition) has thus been concerned to pursue the radical interpretation strategy. But are dualism (in which a hidden and separate meaning is inferred *behind* the non-semantic surface) and a reductive materialist view (in which it is discerned *in* the patterns of the non-semantic) the only options? What if we abandon the assumption common to the materialist accounts and dualism, namely that meaning is not observable, while retaining dualism's commitment to non-reductionism? This gives us a view on which meaning is a public and observable property of certain sounds, marks, or movements, but a non-physical one. So the concept of meaning is a descriptive and factual one, and also, very importantly, a basic observational one. But it is not part of that predominantly quantitative and value-free conceptual scheme we have built up for describing, predicting, and explaining the behavior of inanimate objects; rather, it

belongs to a different but equally fundamental area of thinking, namely the one we use in our relations with other persons. This line of thought is favored by those with Wittgensteinian sympathies. If we accept this view it is likely that the idea of the imagined starting point for radical interpretation, a starting point in which a person knows plenty of non-semantic facts but no semantic facts at all, will come to seem incoherent. The starting point for any thinking is one in which we are observationally aware of the world as containing both semantic and non-semantic facts.

This option will seem to many extremely wild, because it clashes with certain widespread but often unarticulated assumptions about fundamental matters like fact, truth, and perception; so to make it seem even coherent, let alone plausible, we would need to reappraise our views on these topics. But before considering whether we need to embark on that unsettling enterprise, we should surely see whether we cannot find something more immediately congenial in the radical-interpretation camp.

We have then at least the following questions to ask. Is radical interpretation possible? If it is, what are the strengths and weaknesses of the particular variants proposed? If it is not, why does it fail and what should we put in its place?

On this last question, let us remember that in addition to dualism and the Wittgensteinian option sketched we have at our disposal also such views as eliminativism and instrumentalism. The first of these says that psychological concepts, including those of content and meaning, are so confused and/or scientifically ill-grounded that nothing answers to them, and hence no theory of their (true) applicability is required (Churchland, 1988; Stich, 1983). The second says that talk of meaning and content is a useful tool but not to be factually interpreted (Dennett, 1979; 1987).

For various reasons, our entry point into these issues will be consideration of the views of Davidson. The topic we are considering got its name from him, and some important themes emerged in his writings. Focusing on them allows us to identify interesting points of contrast between different theories of radical interpretation. But it is also arguably the case that his view, if correct, contains the seeds of its own destruction, in that it leads to unacceptable claims about the indeterminacy of meaning, and might thus lead us to question the validity of the whole radical-interpretation project.

It is, however, difficult to see at first reading what is of central importance in Davidson's work, because the form of his proposal is substantially but unhelpfully influenced by its history. We turn therefore in §2 to a brief sketch of its development and summary of its mature form.

2 From "Truth and Meaning" to "Radical Interpretation"

A distinction it is useful to have clear is that between providing an analysis of a concept, an account which clarifies its links with other concepts and hence its metaphysics, and providing what I shall call a 'calculus' for that concept, a set of rules for working out if it is applicable to an item on the basis of information about the composition of that item. (See Heal, 1978, for more on this distinction.) For example, consider the claim that sodium nitrite is poisonous to octopuses. A person can show that she knows very well what this means – by unpacking the claim in terms of what will happen to octopuses who ingest sodium nitrite. But this person may be in no position to say whether the claim is true or not, because she has no sodium nitrite or no octopuses, or cannot get the latter to eat the former. Another

person might, by contrast, be in a position to rule on the truth of the claim, because he is in possession of a set of instructions for calculating to what creatures, if any, a chemical compound is poisonous, from a canonical specification of the elements in the compound and their mode of bonding. This person might, however, not fully understand the claim, because for him 'poisonous' is little more than a dummy predicate and he has no grip on its relations to eating, illness, and so forth.

For many concepts, as for 'poisonous,' analysis is independent of calculus. It is a contingent matter whether or not there is a calculus of poisonousness. And even if there is a calculus one can understand 'poisonous' very well without suspecting this, let alone knowing what it is. But for meaning things are different. And this is the starting point for Davidson's discussion in "Truth and meaning" (Davidson, 1984, essay 2; see also Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S). He is concerned to emphasize that an item cannot have sentential-type meaning, that is, be the vehicle for a complete linguistic move, unless it is complex. Any such sentence must be built up from words which, together with their arrangement, determine the meaning of the whole.

His arguments rest heavily upon the fact that natural languages allow for the construction of indefinitely many new sentences. In them a finite number of words are built together in increasingly complicated arrangements by application (repeated if need be) of a finite stock of constructions. Unless we see sentences as built in this way it is entirely mysterious how finite creatures like ourselves could have the language-speaking and -understanding capacities which we do.

These considerations are weighty; and we can cite others which point in the same direction. Even if a language had only a finite number of sentences there would, I suggest, be reasons for thinking that sentence-style meaning requires the existence of words or word-like complexity in the sentences. It is central to the notion of meaning (for contingent *a posteriori* sentences at least) that meaning is one thing and truth-value another. Imagine now some item which has a meaning of this kind and is also false – for example, an item which means 'snow is green.' Try to suppose also that this item entirely lacks semantic complexity in its properties or relations; its having the meaning it does is a one-off matter, and not bound up with any systematic connection with any other meaning-bearing items. We seem here to have something completely unintelligible. The difficulty is not merely epistemological – how could we tell that the item had this meaning? – but also constitutive: what could there be about it which fixes that it is about snow or about greenness? One-word indexical sentences – like 'Fire!' – do not provide a counter-example to this claim. What is required for a linguistic move is a token, and each token of this sentence does exhibit two separable features, namely its type and its spatio-temporal location, which contribute independently to fixing the claim made by the whole.

So Davidson's starting point in "Truth and meaning" is the fact that 'S means that p,' when unpacked, turns out to commit us to the idea that S is a complex item which will have a number of different features, each with its own semantic role, which jointly determine the meaning of S. Let us go along with him in assuming that the features are the presence of identifiable words. This is not obligatory. The features could be things such as color, size, or shape, or they could be relational properties. But the word-hypothesis simplifies exposition without distorting the issues we are concerned with here.

But if S contains words then it must be (potentially at least) part of a language. There must be (the possibility of) other words which could replace some of the words in S, and so the possibility of other sentences which could be built from different combinations of those

words. So 'S means that p' turns out to have the following implication: S belongs to a language, that is, a system of meaningful items, for which a meaning calculus can be given which supplies 'S means that p' as an output theorem. Davidson, in "Truth and meaning" conceives his task to be that of unpacking further what any such meaning calculus would have to be like.

Suppose that S is 'a is F,' that we know that it means that a is F, and that we are convinced that 'a' and 'is F' are the semantically relevant subparts. An adequate meaning calculus will assign properties to 'a,' to 'is F,' and to concatenation: these property assignments will be (some of) the axioms of the meaning calculus, and they must entail as a theorem:

'a is F' means that a is F.

What might they be like? Well, says Davidson, we already know of one sort of axiom set which will do exactly this job, namely that employed in a Tarski-style theory of truth. (See Tarski, 1956. For useful expositions see Haack, 1978, and Evnine, 1991.) All this might merely lead to the thought that a Tarskian calculus of truth-conditions and a Davidsonian calculus of meaning will employ the same shape of machinery. It does not yet entitle us to suppose that the notion of meaning can be unpacked into or analyzed in terms of the notion of truth. But this is the bold step which Davidson proposes. He suggests that we abandon Tarski's requirement that the theorems must have on their right-hand sides something with the same meaning as the sentence mentioned on the left, and that we instead demand only that theorems be true biconditionals about truth-conditions. His conjecture is that this, together with the remaining constraint of using finite semantic structure to generate a theorem about every sentence, must narrow down the number of acceptable calculi to such an extent that anything which satisfies the constraints will, in fact, serve as a calculus of meaning.

The proposal, in summary, is this:

'S means that p' unpacks into
'S is part of a language and a correct truth-conditions calculus for this language can be set up which delivers "S is true iff p" as a theorem.'

(Let us enter a caveat here. Davidson denies (1976, p. 35) that this was his proposal, and he implies that he all along intended further constraints. But however this may be, the line of interpretation I have sketched is the most natural reading of "Truth and meaning," and is what most of its readers took Davidson to be saying.)

The proposal is one about the conceptual connection between the notion of meaning on the one hand and the notions of word, language, structure, calculus, and truth on the other. It claims to offer insight into meaning, but not by invoking anything to do with radical interpretation. Why, then, have we spent so much time on it? Because Davidson's later view is, in part, an attempt to rescue the earlier theory; and the later view retains, unnecessarily to my mind, certain features of the earlier one. These are features which many have found particularly implausible, and they have hindered appreciation of other, more important new elements introduced in the later theory. So unless we know a little of the history we shall not understand properly what is going on.

One damning objection to the early proposal which soon became apparent is this (Foster, 1976; see also Chapter 2,

TO DAVIDSON'S). An acceptable truth-conditions calculus for a language can generate the true theorem that S is true iff p, when it is quite clear that S does *not* mean that p. For example, if the properties P and Q by chance characterize all and only the same things, then a calculus of truth-conditions which assigns Q to a predicate 'F' will do just as well as one which assigns P. Yet looking at the speakers of the language, it may be far more plausible that they are speaking about Q than about P; P might be something these speakers could not be aware of, such as a microstructural property, or it might be some gerrymandered compound property like 'having Q while $2 + 2 = 4$,' which is, of course, coextensive with 'having Q.'

Davidson's moves in "Radical interpretation" provide a response to this. He there offers a proposal about meaning which retains a central element of the earlier work, but locates it in an importantly different setting. The retained element is the idea that to say 'S means that p' is to say that S belongs to a language for which an empirically acceptable calculus of truth-conditions delivers 'S is true iff p' as a theorem. But the empirical conditions which make a calculus acceptable are spelt out differently. The demand now is not that the calculus supply true biconditionals about truth-conditions; it is that it deliver theorems which lead to the speaker of the language satisfying the *Principle of Charity*. We shall return to this in a later section. But what it says is roughly the following: *if* we use our candidate calculus to interpret our subject – that is, whenever the calculus says something of the form 'S is true iff p' and the subject says S, then we take him to be expressing the belief that p – *then* he comes out as by and large a sensible fellow, with a mainly correct view of the world.

This condition may now enable us to distinguish between the two calculi (the one which linked 'F' with property P and the one which linked it with property Q) that were mentioned earlier. It may be that the overall psychology attributed to the subject using Q reveals him as much more coherent and intelligible than the one using P.

We could, then, just add the Principle of Charity to the earlier empirical condition (that the calculus deliver a correct account of truth-conditions) and leave the overall proposal otherwise untouched. It may seem as if this is what Davidson does. And indeed, his later proposal could be summarized in the form of words we used earlier, with the addition of the extra clause.

But it is important to realize that, at the same time as the Principle of Charity comes on the scene, another change takes place in Davidson's thought, namely a shift of attention from the case of thinking about one's own language to that of thinking about a strange language. The home-language case drops away out of explicit consideration. When mentioned, it is presented as just another case of radical interpretation, but one where familiarity disguises this fact from us. The effect of this is that the way of understanding what it is for something to be a correct calculus of truth-conditions which was earlier called upon is no longer available. When we are thinking about the home language we can presuppose an implicit understanding of sentences. This makes it straightforward to test theorems delivered by a candidate theory. Suppose I am already a speaker of English, and I am asked whether '"Snow is white" is true iff rubies contain carbon' is a true biconditional. I may not know the answer instantly, but at least I know what I have to do to find out. But when we change the case to an unknown language, matters are very different. Suppose the candidate theory delivers '"Skuppit gromper" is true iff rubies contain carbon.' How am I to tell whether this is a true biconditional about truth-conditions?

It is easy not to notice how big the shift is between the position of "Truth and meaning" and that of "Radical interpretation." The fact that a very similar formulation can be offered may suggest that all that has happened is that Charity has been added in as an extra

constraint. But a crucial point is that with the change of focus from the home case to the alien, the earlier account of empirical correctness needs replacement. And with its replacement, the whole shape of the enterprise changes. The earlier project aims for an account of linguistic meaning in terms of truth. It tries to illuminate the intensional semantic notion of meaning by unpacking its links with the extensional notion of truth, together with the notions of structure and calculus. This project works within a circle of semantic concepts and it applies only to linguistic items; views about the nature of persons and psychological states are not brought in at all. The second project is, by contrast, in many ways more ambitious. It offers to give an account of *both* meaning *and* truth, which are now put back together as an inextricable pair, neither of which is more fundamental than the other. And the account it offers is extended to apply to psychological states as well as to linguistic items. The story it offers, to put matters extremely briefly, sees meaning, language, truth, and thought all as needing to be explained in terms of patterns discernible in uninterpreted behavior.

Let me now spell out in more detail the fully fledged, later proposal. First some preliminaries. It is supposed that we can identify prior to interpretation the class of sentences which the subject holds true, that is, those he or she is willing to assert sincerely. Patterns of recurrent elements can be discerned in those sentences. The patterns are of such a kind that a finite calculus of truth-conditions can be set up which covers all sentences of the language. This is to say that we can hit on a finite number of axioms which assign semantic properties to the recurrent elements in such a way that we can derive for every sentence some theorem assigning it truth-conditions. The pattern of actions and utterances is also such that use of this calculus as a tool for attributing attitudes makes the subject come out as satisfying the Principle of Charity. Let us call such a calculus 'an empirically correct theory of truth' for the language in question. Now we can state the proposal thus:

> 'S means that p' unpacks as
> 'S is true iff p' is a theorem of an empirically correct theory of truth for the language containing S.

And

> 'A thinks that p' unpacks as
> There is some S which A holds true and S means that p.

Davidson draws the following conclusions from this account: that meaning is a normative notion, implicated with what he elsewhere calls "the constitutive ideal of rationality" (1980, p. 223) and that meaning is to some extent indeterminate, in that alternative calculi of truth-conditions could well be set up which satisfied all the constraints equally well. We shall consider the first of these claims in §4 and the second in §5. But we turn first to another matter.

3 The Basis for Radical Interpretation

How does this approach compare with other theories which share the basic assumption of the radical interpretation program, namely that meaning must somehow be based in or derived from the non-meaningful?

Most other approaches of such character have in common a broadly functionalist and naturalistic orientation. They suppose that the nature of psychological notions is to be elucidated by pointing to their causal role *vis-à-vis* behavior. They thus tend to share a sympathy with the idea of unpacking semantic relations in terms of causal relations. For a person to be thinking that a is F is for him or her to be in a state with suitable causal relations to a and to F-ness. Where they differ is in the spelling-out of what 'suitable' means. Some call on the idea of a language of thought, in which individual words have semantics fixed by their individual causal histories (Field, 1978; Fodor, 1975; 1987). Some give an important role to the etiology of internal structures in natural selection (Millikan, 1984; Papineau, 1987). Others call upon the notion of information (Dretske, 1981). Here I cannot hope to do justice to the variety and ingenuity of theories proposed (see Chapter 8, A GUIDE TO NATURALIZING SEMANTICS). Instead I shall ask where they, or large subgroups of them, seem likely to come into conflict with the Davidsonian view sketched at the end of §2, in the hope that this will bring into focus at least some of the interesting issues in the area.

One immediate point of contrast stands out, namely that most theories other than Davidson's would take as their base not a set of facts about 'holding true,' but the totality of physical facts. In defense of his starting point Davidson writes, "[Holding true] is an attitude an interpreter may plausibly be taken to be able to identify before he can interpret, since he may know that a person intends to express a truth in uttering a sentence without having any idea *what* truth" (1984, p. 135). The general claim here is, however, implausible, and the reason given for it (that it is *sometimes* possible to know that a person intends to express a truth without knowing which) does little to support it. A theory of radical interpretation should be applicable to giant octopuses or super beings emerging from their spaceships as well as to newly encountered human beings. But consideration of such non-human cases makes clear that identification of something as a holding true requires ability to distinguish voluntary from involuntary and linguistic from non-linguistic behavior. There is every reason to suppose that making these distinctions will involve simultaneously making rich hypotheses about the contents of beliefs and purposes. A similar point can be made in connection with speech acts other than sincere assertion, such as commands, stories, and irony (1984, p. 135). Davidson does mention these, but he underplays the fact that distinguishing them from sincere assertions (which we shall have to do if we are to identify the holdings true before interpretation) also requires attributions of rich and complex intentions.

So, in brief, Davidson's proposed radical interpretation starts in a place which is either not available or is not radical. Moreover, even if that place were available, his methodology would have us ignore evidence about the placement of speech in the context of non-linguistic action. But such placement might surely give us useful clues to meaning. And indeed Davidson himself, at another point (1984, p. 162), remarks that a theory of interpretation cannot stand alone, but will need to be integrated within a more comprehensive theory of thought and action.

Why, then, should Davidson have chosen as basic this unsatisfactory 'holding true' notion? One explanation is the powerful influence on him of the Quinean radical translation model. (See Chapter 26, INDETERMINACY OF TRANSLATION.) But another is the desire to carry forward as much as possible of the shape of the "Truth and meaning" theory. There we assumed that we started with an identified body of sentences for which we know truth-conditions, namely the sentences of our own language. The test of a proposed calculus was then simply whether it delivered correct statements about truth-conditions for these

sentences. The nearest we can get to this position, if we shift to a radical situation, is to imagine, first, that linguistic behavior at the start neatly differentiates itself from the rest, and second, that we can establish by observation the conditions under which each sentence is held true. On those assumptions (and *very* importantly, given *also* that holding-true conditions for the most part coincide with actual truth-conditions: see the next section) then we can proceed very much as we did on the "Truth and meaning" story. But the moral of the discussion of non-human cases is that this is not a persuasive line. A desire to duplicate the earlier structure as far as possible has led to the implausible idea that sentences held true can be isolated from the rest of behavior prior to interpretation. The totality of physical facts would be a much less tendentious place to start.

Davidson's proposal, then, needs modification in at least two ways. The first is the change in the imagined starting point, and the second is the need to take account of the relation of linguistic behavior to its context in action. The upshot of such changes would be a theory rather like that proposed by Lewis (1983). The idea is, briefly, this: to say 'A thinks that p' or 'A's utterance S means that p' is correct provided that these claims would be delivered by an acceptable overall theory of A's behavior; a theory is acceptable if it attributes beliefs, desires, intentions, and meanings to A (and thus licenses redescription of some mere movements, noise-makings, and so on as actions and utterances) in such a way that (1) A has a language with a finitely specifiable and reasonably simple semantic structure; (2) A has, by and large, a true, rational, and epistemologically defensible view of the world (that is, he satisfies the Principle of Charity); and (3) A comes out as doing and refraining in action (including linguistic action) in the way expected of a rational person with the intentional states attributed.

What is there in this which might still provoke objections? The Davidsonian idea that there is a link between meaning on the one hand and system or structure on the other is still very much in play. The proposal assumes that nothing can be seen as a meaning-bearer (whether an action expressive of intention or an utterance expressive of belief) unless it is part of a repertoire of other possible items with which it is contrasted. Methodologically, we are exhorted to use these contrasts and the circumstances of their occurrence as key diagnostic elements for interpretation. And this, the insight stressed at the start of §2, is not, it seems to me, something with which any theory of meaning need quarrel. (However, the most unlikely quarrels are pursued. See Fodor and Lepore, 1992, especially ch. 3.)

A potentially more controversial feature of the proposal is that it has a behaviorist flavor. What licenses and makes true attributions of meaning is a certain sort of patternedness in observable behavior. It is not a trivial matter that behavior should be amenable to redescription in terms of a rational psychology, any more than it is a trivial matter that a certain set of intricately shaped and colored flat wooden shapes should fit together in a jigsaw to present a recognizable picture. In both cases it is a matter of individual items locking together, literally or metaphorically, in the right kind of way. And we can imagine sets of items which do not fit. But some might wonder whether such a merely surface feature of behavior was a sufficient condition of thought and meaning, and whether we should not demand also that there be inner causal mechanisms answering in structure to the beliefs and desires attributed.

We shall return to this issue. But first we shall consider another possible source of disquiet, namely the role given in the proposal to the Principle of Charity. Is the Principle, as some have thought, hopelessly parochial? Is it over-optimistic about the likely success of thought? And what does it have to do with some supposed normative aspect to meaning?

4 Interpretation, Charity, Holism, and Norms

To start with, let us look at the early history of the Principle of Charity, when it was, rightly, found unattractive. We should, said Davidson, insist *a priori* that an interpreter find a calculus "which yields, so far as possible, a mapping of sentences held true (or false) by the aliens on to sentences held true or false by the linguist" (1984, p. 27). What this amounts to is that we must make the assumption that pretty well everything we say, and also pretty well everything the aliens say, is true. (Davidson explicitly draws out these anti-skeptical implications of his position in 1984, essay 14. See also Evnine, 1991, ch. 8.)

Why does Davidson make this recommendation? Following our earlier line, a conjecture is that it is because this is the simplest way of converting the "Truth and meaning" proposal for use on an unknown language. For my home language I am able to test whether a truth-conditions calculus is empirically acceptable by seeing whether the theorems it delivers are true. I can do this on a one-by-one basis, using my taken-for-granted knowledge of my own language together with ability to find out about the world. But when we deal with an unknown language, this method is inapplicable. Since I do not understand the sentences, I do not know what to enquire into to test a claim about their truth-conditions. (We have already touched on this point earlier, in §2.) But if I could just *equate* the aliens' being willing or unwilling to assert a sentence with that sentence's being true or false then this lack of understanding need not handicap me. I do not need to investigate the truth of their sentence directly in order to test a candidate calculus; I need only to find out whether the aliens *hold* the sentence true and then, simply assuming that it is true, I find out whether the conditions which the calculus assigns to it do obtain. If they obtain, then the theory is confirmed. Doubtless I must make some allowance for the possibility of the aliens making occasional errors. But if I can insist *a priori* that such mistakes must be extremely rare, then the above simple testing strategy is available.

The early formulation of the Principle of Charity, however, produced objections. It seems to involve the claims, first, that we already have pretty well all the thoughts there are to have (because any set of thoughts can be mapped on to ours), and second, that our beliefs are pretty well 100% correct. These claims appear to underrate the possibilities both for ignorance and for error.

The first claim is in fact not essential to the proposed method. Davidson need not deny our possible ignorance of many aspects of the universe, and he ought to allow that the process of interpretation could be one of substantial learning about the world, that is, it need not merely be one of pairing things we already know with things the alien subject knows. What is more crucial is the claim of substantial correctness in all systems of thought, including our own current one, since it is this claim which underpins the rough equation of 'the subject holds S true' with 'S is true' in our investigative methodology. This second claim is objectionable, however, in that it is either highly implausible or hopelessly unclear. If it is taken to mean that the *actual* sincere utterances of any group of persons must be largely true, it is implausible. It rules out the idea that an extensive part of the lives of a group of persons should be concerned with the pursuit of some chimerical and theoretically ill-based enterprise. How can we rule out *a priori* the idea that our societies might have evolved in such a way as to doom most of us to think and speak, a large proportion of the time, about alchemy, astrology, or historical materialism? But if we shift instead to consider merely *possible* utterances and our willingness to assent to them, the content of the claim becomes extremely unclear, as Davidson himself later hints (1984, p. 136).

Possible utterances form an indefinitely large set, and so quantitative claims about proportions of truth and falsity in the set have no obvious meaning.

So it looks as if "Truth and meaning" is again exerting an unfortunate influence, and has led Davidson to an indefensible position. Before considering whether anything can be rescued, we need to distinguish two different roles for a principle like the Principle of Charity. (Useful moves in this direction are made in Malpas, 1988.) One role is played at the start of the attempt to interpret. Davidson claims that interpretation of words and attribution of beliefs combine to explain utterance. But, he says, we cannot access the beliefs expressed in an utterance prior to and independent of grasp of the meaning of the utterance. Nor can we do the reverse. We need, therefore, to make some initial assumption, in order to begin to test and elaborate any theory. Davidson's way out of the impasse is that we should start by taking it that the beliefs of the other are the same as ours.

There is much that is attractive here. Even those who think that non-linguistic behavior can provide very strong or conclusive evidence for certain beliefs (and hence would reject the detailed Davidsonian proposal, with its stress on language) may recognize an analogous problem in the way in which belief and desire cooperate to produce action. Thus it is a plausible thought that any interpretive enterprise will have to start off by trying out some assumptions in order to break into a circle. And it is also plausible that a good place to start is with the idea that the others are like us.

One alternative to the Principle of Charity, following this line of thought, is sometimes called the Principle of Humanity. It does not recommend us to attribute to our subject as a starting hypothesis the very thoughts we ourselves actually have. But it suggests a close variant, namely the thoughts we would have had, if we had been through someone else's life experiences (Grandy, 1973; Lewis, 1983). This, indeed, looks a useful suggestion and better than straight Charity. If all that was at issue was methodological advice on how to start out interpreting, perhaps we should stop here.

But it is not all that is at issue. However useful as a tip on how to *start*, neither Humanity nor Charity tells us where we shall *end up*. What if our initial assignment of thoughts based on the principle does not work out? What if further investigation makes it seem likely that our subject has different sensory apparatus and/or different emotions and interests from us, and so has significantly different views from any we do or would have had? In these circumstances we must modify our starting hypothesis. And we need at this point to know what constraints there are upon the shape of theory we may move on to, and how to evaluate rival emendations.

Let us now look at one of Davidson's own later remarks. He stresses (as do proponents of the Principle of Humanity) that any interpretation must allow for intelligible error. He writes:

> It is impossible to simplify the considerations that are relevant [i.e., to assessing a proposed theory] for everything we know or believe about the way evidence supports belief can be put to work in deciding where the theory can best allow error and what errors are least disruptive of understanding. The methodology of interpretation is, in this respect, nothing but epistemology seen in the mirror of meaning. (Davidson, 1984, p. 169)

The important theme here is that any set of thoughts we attribute to a subject must come with (or with the possibility of) some intelligible epistemology. This principle, if correct, applies not only to what we guess when we start out interpreting, but to hypotheses at all

stages. Something else stressed by Davidson, and which is part of the same line of thought, is the prominent role given to indexical utterances in providing some anchorage for interpretive theories. We must be able to see some utterances as expressing perceptual judgments about the world around the subject.

What does this insistence on the need for plausible epistemology stand opposed to? Rejection of dualism about meaning is one of the deep-lying assumptions of the enterprise. Classical dualism moves from belief in the *non-reducibility* of facts about psychology (including facts about content and meaning) to claims of their *independence* from the physical. Such independence would allow us to evade the proposed epistemological constraints. We do not, on a dualist view, have to tell a story about how the subject could, so to speak, have got in touch with what we say he is thinking about. We are allowed to suppose that it just is a fact that he is thinking about it. So on a dualist view, content may float free of public facts about a creature's constitution and placement in the world. This is what leads to its constant implicit threat of skepticism, both about other minds and about the external world.

Let us, however, not identify rejection of dualism with acceptance of empiricism. It is one thing to be very unhappy to attribute to a person a thought about some state of affairs which we believe he or she has had no opportunity to perceive, hear of, or theoretically conjecture. It is another to suppose that some sensory reduction can be given for all thoughts. (See Davidson, 1990, for a discussion of the difference.)

Can we say anything else general about sets of thoughts? Yes, says Davidson. We should recognize the so-called holism of meaning. This is the claim that an intentional state with content cannot exist in isolation, but requires the presence in the subject of many other intentional states with suitable contents. For example, anyone who wonders whether the bank is open must believe a fair number of general things, such as that other people exist, that goods are produced and exchanged, and that money exists; and he or she must also have some suitable beliefs about such things as the nature and location of the bank in question.

This view, the holism of meaning, has some resemblance to the claim, argued earlier in §2, of the necessity for any meaning-bearer to be itself complex and to be placed in a system. But they are not the same view. The mere claims of complexity and systematicity carry no immediate implications about the kind of content which is to be carried by the other elements in the system, while the view currently under consideration emphasizes the need for particular kinds of conceptual content to be present. The relation between the two theses deserves more discussion than it can be given here (The word 'holism' is sometimes used also in connection with the view about the joint operation of belief and meaning in giving rise to utterance. But this is a third, and different view. See also Chapter 15, HOLISM.)

The holism of meaning thesis is offered in support of the Principle of Charity, that is, the idea that there must be 'massive agreement' between our thoughts and any other possible set of thoughts and that we must in consequence be substantially right. It is argued that it implies that disagreement cannot exist except against a background of agreement. To redeploy the earlier example, there cannot be disagreement over whether the bank is open unless there is agreement on at least some things, like the existence of other people, the exchange of goods, the existence of money, and so on. So the idea of detecting pervasive error in another thinker (whether we interpret others or they interpret us) is incoherent.

This is the argument. But it does not, in fact, help us with the trouble we had earlier in applying quantitative notions like 'massive agreement' to indefinitely large sets of beliefs; and hence it does not help in seeing exactly what the content of the claim is. Without

detailed supplementary information (about how much agreement is required to underpin one disagreement, about whether one agreed set of beliefs can underpin more than one disagreement, and so forth) nothing follows about quantities. Such supplementation is unlikely to be forthcoming. We would do well, then, simply to jettison the quantitative style of claim.

What, if anything, then remains of the Principle of Charity? Let us first re-express the original insight of meaning holism. It is that for an item to bear a certain meaning it is required that it exist in a setting of other items bearing related meanings. Thus it is only when a rationally coherent and related group of items exists, a group which can be taken to represent some extended portion of a world, that we can attribute content to any member of it.

Let us pause to emphasize something important at this point. Nothing that has been said requires us to insist that a particular thought requires some given, fixed set of other thoughts. We may do so, if we are sympathetic to the idea of analytic truth. But we could instead demand something weaker, namely only that there be *some* suitable setting to anchor a given content. As an analogy, consider what is required for a dot in a cartoon-style picture to represent a living human eye. One might think (adopting the analytic model) that there had to be a dot for the other eye, a line for the mouth, a circle for the face, or at least a reasonable subset of these. But this is not correct. We can imagine a dot which represents the eye of someone peering through a hole in a screen, where none of the rest of the body is depicted. What is required is that there is enough detail in the rest of the picture to make clear that this is what is happening. Similarly for content, one could speculate. We can attribute a given belief in the absence of its usual accompaniments in us, provided there is some other suitable setting to anchor its meaning. And perhaps there are no ways of cataloguing and systematically studying what 'suitable settings' there could be.

So to return to our current question of whether anything useful can be salvaged from the Principle of Charity. We have so far expressed the holism of meaning in terms which suggest that it is relevant only to small subsets within a set of thoughts. But the implications of the holism are wider. The thoughts which are, so to speak, at the edge of one set will need to be in the middle of another. And thus the idea of the 'suitable setting' will spread to encompass the totality of thoughts. The upshot is that the set must, as a whole, be more or less coherent and so represent what could be one world. We cannot have a fragment here and another contradictory fragment there without losing grip on the idea that it is one mind, one subject's point of view, that we are capturing.

If we now combine this thought with the earlier anti-dualist one, we arrive at the following view. Any set of thoughts we attribute to a subject must be a recognizable representation (however incomplete or distorted) of this one world which we share with him or her. It must be something with which we can to some degree sympathize, in that we can see it as the outcome of rationality, that is, cognitive competence in a broad sense, trying to get to grips with this world. This muted version of Charity may be a far cry from a claim of massive agreement and general correctness, but it is not negligible.

Having thus thrown out the early version of the Principle of Charity and summarized what seems important in the later version, let us now turn back to the other question we have been pursuing and ask whether there is anything in the Principle, as now understood, to raise objections from those who hold other, for example functionalist, views about radical interpretation.

The anti-dualist strand, with its emphasis on the 'externalist' idea that the content of a meaningful item is to be, in part at least, fixed by its context (for example, external causal links),

is thoroughly congenial. (For more on the many varieties of externalism see McGinn, 1989.) Some theorists explicitly reject the idea of meaning holism (see, for example, Fodor, 1987, ch. 3; Fodor and Lepore, 1992). But most functionalists would have no difficulty in endorsing this view also. The central functionalist idea is that psychological terms are elements in some folk proto-theory for explaining behavior; their meaning is explained by reference to their explanatory roles in this theory *vis-à-vis* observable behavior and, importantly, *vis-à-vis* each other. The holist idea of linking content to place in some suitable pattern seems a natural development of this. For such a functionalist, just as for Davidson, rationality, the existence of appropriate inferential and content links, is built into the very nature of the psychological. (See Evnine, 1991, pp. 111–112.)

Where, then, could any clash arise? The crucial question is this. Can the key notion of rationality be given at some level a naturalistic, such as a physical, unpacking; or is it centrally a normative notion of a kind which resists such capture? It would be generally agreed that it is normative; it has to do with thinking as one ought, that is, in such a way as to promote the goals of thought. But this does not preclude the existence of a non-normative equivalent. To use a familiar old example, a knife's being good may, given the acknowledged purpose of knives and the causal facts about the world, amount to its having certain physical properties (of weight, shape, sharpness, and so on). Is there something describable in the language of the natural sciences which in a similar sense is what rationality 'amounts to'?

Some may think that we already have such a thing, and that it is given by an amalgam of the (syntactic) rules of inference provided by (a favored) deductive logic, inductive logic, and decision theory. This, however, seems over-optimistic, given the disputes, paradoxes, and unclarities in these subjects. We also need to remember that acquisition of knowledge about the universe does not always take the form of adding empirical information within a fixed conceptual and logical structure, but may involve modification of concepts and hence of patterns of inter-judgmental linkages (consider non-Euclidean geometry, Einsteinian revelations on space and time, and so on). We are not in a position to say that there are no more conceptual upsets ahead; so we cannot now plausibly claim that we know, definitively and completely, what rationality amounts to.

But all the same, is there such a thing as what it amounts to? This is a very close relative of the question that arose above as to whether any notion of analyticity is defensible (see Chapter 23, ANALYTICITY) and, connectedly, of whether there is some limit to the 'suitable settings' which meaning holism requires. Obtaining a reasoned answer to these questions could well require us to have a view on such things as whether there is one complete and final truth about the universe and one set of concepts which that truth demands for its expression, whether there is any hope of human beings attaining that truth, whether (or in what sense) we could be rational if we were constitutionally debarred by the structure of our minds from attaining it, and what the connection is between rationality in theoretical matters (arriving at the truth) and rationality in practical matters (arriving at good decisions). Those who favor an ultimately physicalist account of rationality will incline to assume something optimistic about the possibilities of disentangling theoretical from practical reason and arriving at the complete truth about the universe, while others will be skeptical on these matters. Some skeptics, like Davidson, seem to think that it is in principle impossible that there should be a physicalistic unpacking. Other skeptics, like the instrumentalist Dennett, seem to think that there is such an unpacking, but that our finitude constitutionally debars us from realizing it and consequently our 'rationality' and 'thought' are merely useful fictions.

We can return here briefly to tie up one loose end, namely why Davidson's proposal has, as remarked towards the end of §3, a behaviorist flavor. We can see now that it is not an attractive option for one who is skeptical about a naturalistic unpacking of 'rational' to appeal to internal structures. If there was a fixed structure to rationality then there would be some fixed set of thoughts associated by meaning holism with any given thought, namely that set which the one and only true rationality requires it to be linked with. And it would be natural to insist that a subject who has the thought must contain some physical realization of (at least some part of) the relevant set and its supporting causal linkages. So, on this view, internal mechanisms would be important. But, on the other hand, if there is no fixed structure of rationality, then it is unclear what importance internal mechanisms could have. The only thing that can matter is whether the totality of utterances and behavior can be redescribed in terms of some one of the (perhaps indefinitely many and exceedingly various) rational psychologies. Any investigation into the inner mechanisms which subserve the intelligible behavior must follow after the attempt to make sense of that behavior. The presence of inner mechanisms cannot be used as a separate and prior constraint on whether a subject possesses a certain thought. (Compare here Davidson's rejection of psycho-physical laws in 1980, essay 11.)

5 Indeterminacy of Meaning, Holism, and Molecularity

We have so far been presenting most of our theories, Davidson's included, as though they were straightforward accounts of what non-semantic facts constituted some semantic fact. But we must now grapple with a final complication, namely that the Davidsonion approach seems to imply indeterminacy (1984, pp. 100–101, 153–154). On his view, to say that a noise has a certain meaning is to say that a subject's behavior can be systematized by a certain kind of theory. But the crucial problem is that there is no guarantee that there is only one way of systematizing a given body of behavior. The assignment of meanings to individual moves or noises is merely a record of the fact that the behavior as a whole can be arranged in a certain sort of pattern. But the fact that the pieces of behavior can be arranged in one pattern cannot rule out the possibility of arranging them also in another one. Consider as an analogy the pieces of a jigsaw. We are used to unique-arrangement jigsaws. But there is no conceptual necessity to this, and an ingenious toy-maker could manufacture (perhaps already has manufactured) multiple-arrangement ones. To label an utterance 'a saying that it is raining' is, on the Davidsonian story, like labeling a jigsaw piece 'a mountain summit piece.' In other words, such a labeling tells us that there is at least one overall satisfactory arrangement in which the utterance or piece could play the designated role. But, to re-emphasize, that is no conceptual bar to the existence of another arrangement in which it plays another role. Indeed Davidson goes further and endorses an argument designed to prove that if there is one adequate linguistic interpretation then there must be alternative satisfactory interpretations. His argument (1984, essay 16) is a close relative of some given by Quine (1960, ch. 2; 1969, essay 2). (See also Chapter 26, INDETERMINACY OF TRANSLATION, and Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM; Dummett, 1975; Heal, 1989.)

We could embrace this conclusion and the anti-realist implications which it has for the notion of meaning. But many philosophers find this conclusion intuitively incredible (such as Lewis, 1983, p. 118). We may note two reasons at least for being unhappy with it. The first

is that it seems to threaten the possibility of a realist metaphysical stance in general. (For more detailed arguments to this effect see Heal, 1989, ch. 6.) The second reason is that it undermines our central notions of deliberation and reason-giving. Let me expand on this a little.

We take it for granted, pre-philosophically, that we are capable of arriving at new thoughts by rational inference from our current thoughts, for example by working out what further things are certainly or probably true of the world, given what we already know. To use the jigsaw model, we think of ourselves as, in part, self-building jigsaws, where gaps get filled in or new pieces are added round the edge in the light of the parts of the scene already pictured. But on Davidson's story existing pieces do not have determinate content of the kind which would enable them to be the rational bases for such extensions. On his view, the contents they are to be assigned await determination in the light of the later pieces. The patterns to which we must look in assigning meaning are patterns spread out across time, including the future, and not merely across space. Thus features of later utterances, features which fix what overall patterns they and the earlier ones can form, are partly constitutive of the earlier utterances having the meanings they do. To put matters very picturesquely, the meaning of an individual utterance is not fully present in it at the time when it occurs, but exists, in part, in the future. What goes on now is such that it could have one meaning in the light of one line of future development, and would have another given some alternative.

Now suppose that we want an explanation of the appearance of some later pieces and we want an explanation which is causal or quasi-causal, in the sense that it has to do with the development of some process through time and explains later stages by citing conditions wholly present at earlier times. It is a consequence of the Davidsonian view (as sketched immediately above) that such an explanation cannot invoke the notion of meaning, because the meaning is not, on that story, present at the time of the utterance to do any causing. Common sense, however, is strongly committed to the possibility of such meaning-invoking and quasi-causal explanation. It is thus committed to what Dummett calls 'molecularity' in a theory of meaning – that is, the idea that grasp of individual concepts (and relatedly the having of determinate individual thoughts into which those concepts are assembled) has some real ontological and explanatory priority *vis-à-vis* the total assemblage of thoughts at which a person arrives (Dummett, 1975). Davidson's holism precisely denies this. So, somewhat paradoxically, Davidson's strong insistence on the rationality of thought as it is spread out through time threatens to deprive us of any dynamic rationality in determination of our futures.

It is the commonsense conviction of molecularity which gives such power and attractiveness to those functionalist theories which emphasize the causal role and determinate nature of inner structures. So one strategy for avoiding the unwelcome indeterminacy would pursue those questions about the nature of rationality which were mentioned at the end of the last section. Such a strategy would also emphasize the importance of determinate causal connections between aspects of the world and states of persons in fixing semantic relations. The hope would be that some acceptable naturalistic theory, using these kinds of materials, can be built. (For more on semantic naturalism, see Chapter 8, A GUIDE TO NATURALIZING SEMANTICS.)

But there is a different move which has been proposed by philosophers influenced by Wittgenstein and skeptical of the success of the naturalistic program (see McDowell 1981; 1982; 1984; Heal, 1989; Mulhall, 1990). They propose rejection of the initial assumption made in §1, which grounded the whole search for a theory of radical interpretation, that

possession of meaning had to be an inferred or constructed state of affairs and is such that it cannot be simply observed. Let us turn to look at this more closely.

It is surely true that we could be in a situation of needing to engage in something like 'radical interpretation.' That is to say, we could be (a) confronted with some complex moving physical object, which might plausibly be taken to be a thinking, talking creature but (b) not yet certain that the creature is indeed a thinker or talker. In such a case there is nothing for it but to assemble what information we can about the creature's behavior and circumstances – information which can be stated without commitment on what it thinks and means, because *ex hypothesi* we do not as yet have firm views about this – and to see what we can conjecture.

But it does not follow immediately from this that meaning is always and essentially a theoretical or inferred matter. Consider a parallel case. I may be confronted by an array of colors and shapes, but unclear as to whether I am seeing a material object. I may then try to assemble facts about the nature of the array and how it changes under various circumstances, in order to help me decide. These facts will be statable without commitment to a view about whether I am seeing an object and/or its nature. We need not conclude, however, that sense-datum theories in their classic form should be resurrected, and that material-object statements should be seen as inferred from or constructed out of sense-datum ones. One crucial idea in seeing that it need not follow is that the investigation of sensory experience will often itself presuppose facts about material objects and our uncontroversial perception of them. For example, it may take the form of seeing what happens *when I put on my spectacles* or *when I move my head*. It is extremely doubtful that we can, in turn, take these conditions to be inferred from or constructed from sense data: and if they cannot, what we arguably have is a conceptual scheme in which both material-object and perceptual-experience judgments are, in different and interlocking ways, 'observational,' and in which neither can be regarded as more fundamental than the other.

Could it be the case that some analogous possibility holds for meaning? On such a view, claims about physical items, with their causes, circumstances, patterns, and so on, would have conceptual links with claims about meanings (just as claims about perceptual experience have conceptual links with claims about material objects), but the latter would not be reducible to the former. And both would be fundamental and observational elements of our conceptual repertoire. On such an approach, it might be possible to combine respect for the normative and holistic elements Davidson stresses with the molecularity and determinacy to which common sense is committed (see McDowell 1981; 1982; 1984; 1986; Heal, 1989).

It is, however, becoming clearer than ever that our questions about the nature of meaning and the possibility of radical interpretation are linked with other fundamental philosophical questions. Of the views sketched in this chapter, different ones will seem attractive, depending upon one's sympathies on certain basic matters; and adjudication between the options discussed is possible only in the light of the persuasiveness of large-scale philosophical positions.

References

Churchland, P. M. 1988. *Matter and Consciousness*. Cambridge, MA: MIT Press.
Davidson, D. 1976. "Reply to Foster." In Evans and McDowell, 1976, pp. 33–41.
Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.
Davidson, D. 1990. "Meaning, truth and evidence." In *Perspectives on Quine*, edited by R. Barrett and R. Gibson, pp. 68–79. Oxford: Blackwell.
Dennett, D. 1979. *Brainstorms*. Hassocks, Sussex: Harvester Press.
Dennett, D. 1987. *The Intentional Stance*. Cambridge. MA: MIT Press.
Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Dummett, M. 1975. "What is a theory of meaning?" In *Mind & Language*, edited by S. Guttenplan, pp. 92–122. Oxford: Clarendon Press.
Evans, G. A., and J. McDowell, eds. 1976. *Truth and Meaning*. Oxford: Clarendon Press.
Evnine, S. 1991. *Donald Davidson*. Cambridge: Polity Press.
Field, H. 1978. "Mental representation." *Erkenntnis*, 13: 9–61.
Fodor, J. 1975. *The Language of Thought*. New York: Thomas Y. Crowell.
Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
Fodor, J., and E. Lepore. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.
Foster, J. A. 1976. "Meaning and truth theory." In Evans and McDowell, 1976, pp. 1–32.
Grandy, R. E. 1973. "Reference, meaning and belief." *Journal of Philosophy*, 70(14): 439–452.
Haack, S. 1978. *Philosophy of Logics*. Cambridge: Cambridge University Press.
Heal, J. 1978. "On the phrase 'theory of meaning.'" *Mind*, 87(347): 359–375.
Heal, J. 1989. *Fact and Meaning*. Oxford: Blackwell.
Lewis, D. K. 1983. *Philosophical Papers*, vol. 1. Oxford: Oxford University Press.
Malpas, J. E. 1988. "The nature of interpretive charity." *Dialectica*, 42: 17–36.
McDowell, J. 1981. "Anti-realism and the epistemology of understanding." In *Meaning and Understanding*, 2nd edn, edited by H. Parrett and J. Bouveresse, pp. 225–248. New York: Walter de Gruyter.
McDowell, J. 1982. "Criteria, defeasibility and knowledge." *Proceedings of the British Academy*, 68: 455–479.
McDowell, J. 1984. "Wittgenstein on following a rule." *Synthese*, 58(3): 325–363.
McDowell, J. 1986. "Functionalism and anomalous monism." In *The Philosophy of Donald Davidson: Perspectives on Actions and Events*, edited by E. Lepore and B. McLaughlin, pp. 385–398. Oxford: Blackwell.
McGinn, C. 1989. *Mental Content*. Oxford: Blackwell.
Millikan, R. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.
Mulhall, S. 1990. *On Being in the World*. London: Routledge.
Papineau, D. 1987. *Reality and Representation*. Oxford: Blackwell.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Stich, S. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Tarski, A. 1956. "The concept of truth in formalised languages." In *Logic, Semantics and Metamathematics*, pp. 152–278. Oxford: Clarendon Press.

Postscript

ALEXANDER MILLER

Jane Heal's entry discussed the relationship between Davidson's "Truth and meaning" papers (essays 1–5 of Davidson, 1984) and his "Radical interpretation" papers (essays 9–12 in the same volume). In this postscript, we look briefly at a third set of papers by Davidson in which radical interpretation is central. These might be called Davidson's "anti-convention" papers: we'll look in particular at Davidson (1986) and Davidson (1992).[1]

1. Davidson's leading idea in the "Truth and meaning" and "Radical interpretation" papers was, very roughly, that we can make headway in the philosophical examination of the notion of meaning by constructing formal theories of meaning for natural languages, where these theories take the form of empirically verified Tarski-style theories of truth-conditions constructed in accordance with the Principle of Charity. Clearly, the notion of a natural language as such occupies a central place in this account, and one might expect the idea that speakers master the same language to play a role in explaining how they are able to communicate with each other. Moreover, the rules shared by competent speakers of a language are naturally thought of as *conventions*. For example, in the context of a discussion of Davidsonian theories of meaning, speaking of semantic primitives – expressions assigned proper axioms within Davidsonian theories – Crispin Wright comments:

> I do not think that we can attach any content to the supposition that such an expression has a meaning except insofar as meaning is thought of as constituted, at least in part, by *convention*. (Wright, 1986, p. 211)[2]

Startlingly, Davidson rejects all of these ideas. He writes:

> [T]here is no such thing as a language, not if a language is anything like what many linguists and philosophers have supposed. There is therefore no such thing to be learned, mastered or born with. We must give up the idea of a clearly defined shared structure which language-users acquire and then apply to cases. (Davidson, 1986, p. 107)

He describes the conception of a language he opposes as follows:

> [I]n learning a language, a person acquires the ability to operate in accord with a precise and specifiable set of syntactic and semantic rules; verbal communication depends on speaker and hearer sharing such an ability, and it requires no more than this. (Davidson, 1994, p. 110)

And he rejects the idea that the notion of convention plays an essential role in explaining linguistic communication:

> [W]e should give up the attempt to illuminate how we communicate by appeal to conventions. (Davidson, 1986, p. 107)

2. Davidson considers a range of abilities possessed by competent speakers: abilities to interpret malapropisms, to understand garbled or incomplete utterances, to interpret unfamiliar words or phrases, to correct slips of the tongue, and so on. He argues that such phenomena threaten the idea that linguistic competence consists in command of a language governed by shared rules and conventions that have been learned in advance. In all of these cases, command of a language – so construed – is neither necessary nor sufficient for communication.

Mrs Malaprop intends that her utterance of "This is a nice derangement of epitaphs" be interpreted by me as true if and only if this is a nice arrangement of epithets. Despite having no theory that would allow me to interpret the utterance as she intended – according to the theory I bring to the occasion, it is true if and only if this is a nice derangement of epitaphs[3] – I grasp perfectly well what she intended to say. Mrs. Malaprop and I being party to shared rules and conventions is neither necessary nor sufficient for us to communicate successfully.

Likewise, consider Donnellan and MacKay, two philosophers debating whether the former holds a "Humpty Dumpty" theory according to which a word can mean whatever you want it to mean. At the end of his comments, Donnellan mischievously utters "There's glory for you!" with the intention of being interpreted as saying *there's a nice knockdown argument for you!* Despite the fact that they have no shared theory that assigns the relevant meaning to "glory," MacKay understands perfectly well what Donnellan has said:

> What is common to the cases [Mrs Malaprop and Donnellan] is that the speaker expects to be, and is, interpreted as the speaker intended, although the interpreter did not have a correct theory in advance. (Davidson, 1986, p. 98)

In each case, "the interpreter has adequate clues for the new interpretation," clues that are intentionally provided by Donnellan and unintentionally provided by Mrs Malaprop. In Humpty Dumpty's exchange with Alice, by contrast, Humpty Dumpty fails where Donnellan succeeded, because Humpty Dumpty knows that Alice has no suitable clues to go on.

The alleged irrelevance of shared convention to communication leads Davidson to conclude: "we must pry apart what is literal in language from what is conventional or established" (1986, p. 91). He suggests replacing the association of literal with conventional meaning with what he calls *first meaning*, a notion "that applies to words and sentences as uttered by a particular speaker on a particular occasion" (1986, p. 91). The first meaning of Mrs Malaprop's utterance is *this is a nice arrangement of epithets*, the first meaning of Donnellan's utterance is *there's a nice knockdown argument for you!*, while Humpty Dumpty's utterance has no first meaning.[4]

Davidson suggests that these points generalize: "Mrs Malaprop and Donnellan make the case general. There is no word or construction that cannot be converted to a new use by an ingenious or ignorant speaker" (1986, p. 100). So, in general, communicative success does not depend on speaker and hearer having shared command of rules and conventions learned in advance. If we think of a language as a structure of rules and conventions learned in advance and whose sharing is necessary and sufficient for communication, then, "there is no such thing as a language" (1986, p. 107).

3. Davidson thus rejects the proposition

(A) Shared grasp of rules and conventions is necessary and sufficient for successful linguistic communication.

We'll evaluate Davidson's "malapropism" argument for (A) in §5 below. In this section, we'll briefly consider Davidson's remarks on how his view of these matters relates to those developed in Kripke (1982).

In "The second person" (Davidson, 1992), Davidson represents Kripke's Wittgenstein (KW) as arguing that meaningful language is necessarily social in a sense that is both implausible and stronger than that advocated by Davidson himself. Davidson sees the role that communal agreement plays in KW's "skeptical solution" as adding up to a commitment to (A): KW holds that "speaking a language requires … that more than one person must speak the same language" (Davidson, 1992, p. 114). Moreover, KW "depends on the second person, or a community, to embody a routine which the speaker can share" (Davidson, 1992, p. 121). KW thus embraces (A), a claim allegedly undermined by Davidson's "malapropism" argument.

In attributing (A) to KW, though, Davidson fails to recognize the "skeptical" nature of the solution KW offers to the "skeptical paradox" about rule-following, and as a result he mischaracterizes KW's conception of the sense in which language is necessarily social. Underlying (A) is the idea that shared grasp of rules *explains* successful linguistic communication. This is clearly Davidson's main concern: he writes "The concepts of conventions or rules, like the concept of a language, cannot be called on to justify or explain linguistic behavior" (1992, p. 111). However, this is not something that KW would dispute: since he proposes a "skeptical solution" to the rule-following paradox, KW denies that there are facts about which rules agents are following, and so *a fortiori* denies that there are facts about rules that can play a part in the explanation of linguistic behavior and successful communication. This is clear in Kripke's remarks on the role played by agreement in the skeptical solution:

> On Wittgenstein's conception, [w]e cannot say that we all respond as we do to "68 + 57" *because* we all grasp the concept of addition in the same way … For Wittgenstein, an "explanation" of this kind ignores his treatment of the skeptical paradox and its solution. There is no objective fact – that we all mean addition by "+," or even that an individual does – that explains our agreement in particular cases. (Kripke, 1982, p. 97)

And in his summary of Wittgenstein's view as it struck Kripke, he writes:

> The success of [our rule-following practices] … depends on the brute empirical fact that we agree with each other in our responses [to 'go on' in particular ways]. Given the skeptical argument [for the claim that there are no meaning-facts], this success cannot be explained by the fact that we all grasp the same concepts. (Kripke, 1982, p. 109)

In locating the social nature of KW's view in (A), then, Davidson has gone badly astray.[5]

4. Davidson represents himself as proposing "a weaker and more plausible alternative to Kripke's proposed account of what is required in order to mean something by what one says" (1992, p. 117). As we've seen, this is in part because Davidson erroneously ascribes (A) to KW. But Davidson also fails to see that the preconditions for the possibility of meaning in his own account are of roughly equal strength to those actually imposed by KW.

To see this, consider Davidson's "triangulation" argument concerning the determinacy of concepts. According to Davidson, if we consider a child, say, in isolation, there is no fact of the matter as to whether its uses of the word "table," regularly tokened in the presence of tables, are reactions to, or about, tables as opposed to, say, stimulations of the nerve-endings activated in the presence of tables:

> Why not say that its [the child's] responses are not to tables but to patterns of stimulation at its surfaces, since these patterns of stimulation always produce the response, while tables produce it only under favorable conditions? (Davidson, 1992, p. 118)

Subject to certain conditions, the addition of a second person – an interpreter – to the situation can allow the indeterminacy to be resolved:

> [T]he child finds tables similar; we find tables similar; and we find the child's responses to the presence of tables similar. It now makes sense for us to call the responses of the child responses to tables. Given these three patterns of response we can assign a location to the stimuli that elicit the child's responses. The relevant stimuli are the objects or events we

> naturally find similar (tables) which are correlated with responses of the child we find similar. It is a form of triangulation: one line goes from the child in the direction of the table, one line goes from us in the direction of the table, and the third line goes between us and the child. Where the lines from child to table and us to table converge, "the" stimulus is located. (Davidson, 1992, p. 119)

Moreover:

> [This] kind of triangulation … is necessary if there is to be any answer at all to the question what its concepts are concepts of. If we consider a single creature by itself, its responses, no matter how complex, cannot show that it is reacting to, or thinking about, events a certain distance away rather than, say, on its skin. (Davidson, 1992, p. 119)

The conditions that have to be satisfied in order for triangulation to effect determinacy are that the child and the interpreter have similar natural inclinations to group various worldly items together:

> For [triangulation] to work … *the innate similarity responses of the child and teacher – what they naturally group together – must be much alike*; otherwise the child will respond to what the teacher takes to be similar stimuli in ways the teacher does not find similar. A condition for being a speaker is that there must be others enough like oneself. (Davidson, 1992, p. 120, emphasis added)

Davidson's condition – that there be others with like "innate similarity responses" – is virtually identical to the condition imposed in KW's skeptical solution. In the skeptical solution, the sustainability of the language game in which we are permitted to assert that so-and-so has mastered such-and-such a concept depends on the fact that the individual belongs to a community whose members have the same natural inclinations to, for example, continue the series 2, 4, 6, …, *in the same way* beyond a certain point:

> [O]ur license to say of each other that we mean addition by "+" is part of a "language game" that sustains itself only because of the *brute fact* that we generally agree. (Kripke, 1982, p. 97, emphasis added)

> It is an *empirical fact* that … individuals often are disposed to give responses in concrete cases with complete confidence that proceeding this way is "what was intended" … [T]he success of our practices depends on the *brute empirical fact* that we agree with each other in our responses. (Kripke, 1982, pp. 108–109, emphases added)[6]

Kripke also speaks of "brute inclinations" (1982, p. 112, n. 88) and "unhesitating responses" (1982, p. 86). The fact that members of a community have co-incident primitive inclinations is essentially what Davidson himself is referring to when he speaks of interpreter and interpretee having the same "innate similarity responses." Contrary to what Davidson claims, then, the social conditions he imposes on the possibility of meaning are not weaker than those imposed by Kripke.

5. Returning very briefly to the "malapropism" argument outlined in §2 above, we can raise a doubt as to whether it establishes its intended conclusion even if we grant that Davidson correctly describes the relevant linguistic phenomena. Let's suppose that Davidson

is right when he says that I understand Mrs Malaprop despite the different meanings we attach to "derangement" and "epitaph," and so grant that he is right to claim:

> (a) For any given utterance e that speaker A makes in the presence of speaker B, it is possible that e results in successful communication despite the fact that e contains an expression to which A and B attach a different meaning.

However, what Davidson really needs in order to establish "that there is no such thing as a language" is rather:

> (b) It is possible that: most utterances e made by speaker A in the presence of speaker B result in successful communication despite the fact that e contains an expression to which A and B attach a different meaning.

Moving from (a) to (b) would appear to commit a kind of quantifier shift fallacy: for any given meal I eat, it is possible that I remain healthy despite its consisting solely of a Snickers bar, but this does not imply that it is possible for me to remain healthy on a diet consisting exclusively of chocolate covered peanuts. Likewise, the possibility that some *episodes* of communication are successful despite the absence of shared meaning doesn't entail the possibility of a generally successful communicative *practice* in the absence of shared meaning.

Even if Davidson is right about cases like Mrs Malaprop, therefore, he may not be warranted in moving to the conclusion that successful communication does not require shared meaning, rules, and conventions.[7]

6. In this short postscript we have been able to consider only a couple of the many fascinating issues raised by Davidson's papers on the role played by rules and conventions in linguistic interpretation. A fuller account would need to examine whether his apparatus of "prior" and "passing" theories (Davidson, 1986, p. 101) renders his anti-conventional stance consistent with that of "Truth and meaning." For further discussion, see LePore and Ludwig (2005, ch. 17), and Gluer (2011, ch. 2).[8]

Notes

1 Other important papers in this set include Davidson (1982), Davidson (1989), and Davidson (1994).

2 The similarity between Davidson's ideas on the construction of formal theories of truth-conditions for natural languages and the approach taken in Wittgenstein's *Tractatus-Logico Philosophicus* has often been noted (see, e.g., Kripke 1982, pp. 71–72, n. 60). Interestingly, the *Tractatus* appears to assign a role to convention of a piece with that which Davidson rejects. See, for example, 4.002, where Wittgenstein writes "The tacit conventions on which the understanding of everyday language depends are enormously complex" (as translated in Wittgenstein, 1962).

3 Of course, talk of a speaker "having a theory" is for Davidson simply shorthand for the idea that the theory serves to correctly describe his linguistic competence. See Davidson (1986, pp. 95–96).

4 The emphasis on the notion of first meaning as more fundamental than that of shared language is one major bone of contention in the exchange between Dummett (1986) and Davidson (1994).

5 Davidson confesses to ignoring the "skeptical" aspect of KW's view (1992, p. 113, n. 7), but he fails to see that this badly distorts his account of KW's position on the social nature of meaning and rules.

6 In fact, Kripke tries for something weaker than this: so long as the individual is not *considered in isolation*, he may follow rules even if physically isolated (Kripke, 1982, p. 110). If sustained, this would make Kripke's condition even weaker than Davidson's.

7 Davidson's remarks (1994, p. 119) on the "theoretical possibility of communication without shared practices" seems to me to leave the quantifier shift worry untouched.

8 Thanks to Bob Hale, Jane Heal, Ali Hossein Khani, and Daniel Wee for helpful discussion.

References

Davidson, D. 1982. "Communication and convention." Reprinted in Davidson, 1984, pp. 265–280.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davidson, D. 1986. "A nice derangement of epitaphs." Reprinted in Davidson, 2005, pp. 89–107.

Davidson, D. 1989. "James Joyce and Humpty Dumpty." Reprinted in Davidson, 2005, pp. 143–157.

Davidson, D. 1992. "The second person." Reprinted in Davidson, 2001, pp. 107–121.

Davidson, D. 1994. "The social aspect of language." Reprinted in Davidson, 2005, pp. 109–125.

Davidson, D. 2001. *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press.

Davidson, D. 2005. *Truth, Language, and History*. Oxford: Oxford University Press.

Dummett, M. 1986. "A nice derangement of epitaphs: some comments on Davidson and Hacking." In *Truth and Interpretation*, edited by E. Lepore and B. McLaughlin, pp. 459–476. Oxford: Blackwell.

Gluer, K. 2011. *Donald Davidson: A Short Introduction*. Oxford: Oxford University Press.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

LePore, E., and K. Ludwig. 2005. *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Oxford University Press.

Wittgenstein, L. 1962. *Tractatus Logico-Philosophicus*, translated by D. F. Pears and B. McGuinness. London: Routledge.

Wright, C. 1986. "Theories of meaning and speakers knowledge." Reprinted in *Realism, Meaning and Truth*. Oxford: Blackwell, 2nd edn, 1993.

14

Propositional Attitudes

MARK RICHARD

Propositional Attitudes and Philosophy of Language

What are propositional attitudes? An informal answer is that they are relations – like belief, fear, hope, knowledge, understanding, and assuming – between minds and propositions. A somewhat sharper answer identifies them with the sort of mental state that (normally) has truth-conditions (or the like) in virtue of being representational.[1] We can distinguish among propositional attitudes in terms of their differing connections to behavior, to perception, and to one another.

A variety of objections (which for want of space we won't discuss) might be raised to this characterization. Some have wanted to reserve the term 'propositional attitude' for states which are "in principle accessible" to consciousness, or that are "inferentially integrated" with other propositional attitudes (see Chapter 15, HOLISM, and Chapter 12, TACIT KNOWLEDGE, §2). At issue is the status of the states ascribed to us by theories in linguistics and cognitive science.[2] Some say that some perceptual states (seeing a lion dance, for example) satisfy our definition, but are importantly different from propositional attitudes (because they are relations to events or other concrete entities).[3]

Some of the contention and research surrounding propositional attitudes and sentences ascribing them (call the latter APAs) results from their importance to epistemology, philosophy of mind, and action theory. Why has philosophy of language been so concerned with attitudes and their ascription?

Perhaps the primary reason is the view that propositional attitudes are relations to *propositions*. On many views, propositions both are closely related to meanings (and thus critical to an account of linguistic competence) and are what is in the first instance true or false (and so are integral to an account of how thought and talk relates to the world). On the simplest plausible version of this view,

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

(1) Propositional attitudes are binary relations picked out by attitude verbs such as 'believes' and 'says.'[4]
(2) A use of a (declarative) sentence *expresses* a proposition, save in cases in which the sentence is semantically defective, as perhaps are sentences with empty names.[5]
(3) Propositions have truth-conditions and thus truth-values and modal properties such as necessity. A sentence inherits these from the proposition it expresses.
(4) In a use of an APA *a Vs that S*, V is an attitude verb, *that S* names a proposition expressed by S; the ascription is true provided that what *a* names is related, by the relation *V* picks out, to the proposition.[6]
(5) There is an (axiomatizible) account assigning, to a sentence-type S and a context of use c, the proposition that is expressed by S as used in c. This proposition is what's named in c by *that S*.
(6) A sentence's meaning is given by a rule which enables one to tell, given appropriate non-linguistic information about a context or a use (for example, who is speaking, what time it is), what proposition the sentence expresses. To be a competent speaker is to know such a rule.

The spirit of (1) through (4) informs a good deal of what Frege and Russell had to say about language, truth, and thought. Something like (5) informs the work of those of us who count Tarski and Montague along with Russell and Frege as founders of modern philosophy of language.

Call (1) through (5) the *relational account* of attitudes and their ascription. It is a short step from the relational account to the idea that when a sentence is used (literally) to convey information, the information conveyed is the proposition the sentence expresses: the information conveyed by Sam's literal use of S is naturally identified with both what's said by S and (given the relational account) with the referent of the complement clause in a true use of 'In uttering S, Sam (literally) said that T.' Given that the role of a semantic theory is in large part to give an account of how sentences are used to convey information, it follows that a good deal of semantics will be concerned with explaining the association of objects of propositional attitudes with sentences and their uses.[7]

David Kaplan popularized (6); he calls the sort of rule in question the *character* of a sentence (Kaplan, 1977; see also Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS). It is controversial in ways that (1) through (5) are not: perhaps a person can be a competent speaker of English in virtue of having syntactic knowledge and a variety of abilities which needn't add up to propositional knowledge of the rules assigning propositions to sentence uses.[8]

Some reject the relational account's implication that semantics is about what's said. For example, Davidson identifies theories of meaning with theories giving a Tarskian account of truth.[9] But even the dissenters recognize in one way or another the importance of another problem posed by attitude ascriptions. This is what Max Cresswell has called the *hyperintensionality* of APAs: Expressions which have the same possible-worlds intension – and so, in general, can be substituted for one another outside of APAs and quotational contexts – cannot be so exchanged in APAs. This anomaly constitutes a second reason why APAs are important to the philosophy of language.[10]

This problem might well be divided into two sub-problems. First of all, what otherwise appear to be logical laws apparently fail when applied to APAs.[11] If the argument from A to B is logically valid, then *If A, then A* and *If A, then B* are logically equivalent, no matter how hairy the proof of the latter might be. So

Donald knows that if A, then B

appears to come from

Donald knows that if A, then A

by substitution of logical equivalents. But few think pairs so related must be equivalent in truth-value. The first problem of hyperintensionality, then, is to explain what transformations within the scope of attitude verbs are logically valid.

Notoriously, proper names of the same thing are intersubstitutable within modal contexts *salva veritate*, but are not in the scope of attitude verbs. For example,

Odile thinks that Twain is dead
Odile thinks that Clemens is dead

apparently needn't agree in truth-value. In fact, even expressions, like 'yell' and 'shout,' which appear to be *synonyms* are not intersubstitutable within the scope of attitude verbs. Tyler may take 'yell' and 'shout' to be synonyms, but think that there are people who understand these expressions – and so have beliefs about yelling and shouting – but who have doubts that every shout is a yell. It is quite plausible that in this case

Tyler thinks that some doubt that all who shout yell

may be true, while

Tyler thinks that some doubt that all who shout shout

is not. But the latter comes from the former by substitution of (apparent) synonyms for synonyms.[12]

That even substitution of apparent synonyms fails within attitude ascriptions presents a *prima facie* counter-example to the claim that natural languages have semantics which satisfy a non-trivial principle of compositionality, on which the semantic properties of a complex expression are determined, in a way which can be spelled out in a finite theory, by the semantic properties of its parts and by its syntax. For synonyms, one would have thought, have identical semantic properties. The second problem which the hyperintensionality of attitude ascriptions poses is that of finding an account of the semantic properties of expressions which provides a satisfactory account of the truth-conditions of attitude ascriptions (and other constructions, of course), while also allowing for a compositional account of the semantics of natural languages.

This last problem has proven remarkably difficult. Indeed, the arguable failure of all attempts to arrive at a satisfactory account of APAs moved Stephen Schiffer (1987) to argue that natural languages simply have no compositional semantics, in the sense of an axiomatizable theory which characterizes the conditions under which a use of an arbitrary sentence of the language is true or false.

The balance of this chapter discusses the prospects for a satisfactory account of the semantics of natural language attitude ascriptions.

Questions about Propositions

Assume for the next few sections that to ascribe an attitude with a sentence of the form *a Vs that S* is to say that (what) *a* (names) bears a relation, picked out by *V*, to the proposition determined by the complement of the verb. Much of the work on APAs centers upon the search for an account of propositions that (married to the just mentioned assumption) underwrites our intuitions about the truth-conditions of attitude ascriptions.

There are three sorts of questions one might ask about propositions. First off is a metaphysical one: What sorts of things are they? Among candidate answers are the claims that propositions are: *sui generis* abstract entities; states of affairs – congeries of objects and relations or constructions out of such; properties – for example, the property a state of affairs has when its obtaining necessitates that it is now snowing in Lille; linguistic entities – interpreted sentences or utterances, for example.

The metaphysical question is of particular interest if one thinks that propositions must be "intrinsically representational": not only do they possess truth-conditions (and thereby represent), but their representational properties are not derivative. One might think this if one thinks (as Frege did) that the fact that talk and thought is representational is explained in terms of the representational properties of what one "grasps" in thought and talk. One might also infer the "intrinsic representationality" of propositions from the fact that propositions have their truth-conditions essentially, and thus must have them even in the absence of minds and their activities.

Those who think that propositions are intrinsically representational have offered a variety of accounts that are supposed to explain how there can be such things. Scott Soames suggests that representation arises in cognitive activity, in particular in the predication of properties and relations. To think snow white is to predicate whiteness of snow; propositions, Soames says, are types of such cognitive events, the proposition that snow is white being the event type of predicating whiteness of snow. On this view, propositions are essentially representational since they are the sort of thing that happens when one represents. Jeff King suggests that propositions are (roughly put) objects and properties standing in the kind of (representational) relations that (for example) Renee and walking away stand in, given the existence of sentences like 'Renee walks away.' Other theorists have given kindred accounts of propositions. [13]

One might wonder why propositions need to be "intrinsically representational" to begin with. Those who think that propositions are sets of possible worlds, ways things might be, or something fact-like ("states of affairs") will deny that propositions are representational in anything more than a derivative sense. What is representational is cognitive and linguistic activity; propositions – that is, the sorts of things picked out by complement clauses – are things we use to classify the states of the things that represent.[14]

Second, there is, as we might put it, an alethic question. The (more or less) standard view of attitudes like saying and thinking is that what is said or thought is absolutely true or false – its truth does not vary across time, place, or taste. The alethic question is: Really? Here are reasons to think that the standard view isn't right. (a) David Lewis claims that objects of thought can't be propositions *cum* things absolutely true or false because one who knows every proposition doesn't thereby know everything there is to know. A world might contain two propositionally omniscient gods, Gary and Ed: each knows that Gary lives in Shutesbury and eats corn, that Ed lives in Leveritt and eats peas, that Gary is tall and Ed is not, and so on. They could, Lewis says, know every proposition and thus know exactly what

world they occupied though neither one knew whether *he* ate peas or corn. Lewis's solution to this puzzle is to propose that the objects of assertion and belief are *properties*, not propositions. For Ed to know that he eats peas is to "self-ascribe" the property of eating peas; for Gary (or Ed) to know that Ed eats peas is for him to self-ascribe the property of being such that Ed eats peas. (b) Some say that facts about (how we report) belief identity and retention show that what is said and thought changes truth-value over time. Suppose Jane sees June in 2013 and thinks 'she's pregnant'; Janine sees June in 2015 and thinks 'she's pregnant.' This seems to imply that when Jane saw June she thought June was pregnant, and when Janine saw June that's what she thought/she thought so too. But since Jane might have been right while Janine was wrong, it appears that the thing each thinks can be true at one time, false at another. (c) A more radical objection to the standard view observes that: (i) the standards for being rich (or round or rotund or…) vary across conversations, so that a use of 'June is rich [for a resident of Rome]' may be true in one conversation, false in another, but; (ii) even when we know that standards so vary, we take she who utters 'June is rich' in the first conversation to disagree with someone who in the second conversation denies it. (i) and (ii) seem to imply that the object of (assertion, belief, and) disagreement is something whose truth varies with the standards or presupposition in a conversation.

The literature on (b) and (c) is too vast to take up here.[15] We will return to Lewis's example and the topic of belief *de se* at the end of this chapter.

Semantics and Structure

I said there were three questions one might ask about propositions. The third is one about the way propositions are related to the sentences that express them. Propositions, we are supposing, are picked out by complement clauses, things like the phrase 'that it's raining.' One would like to know exactly how a proposition is determined by a (use of a) complement clause; ideally, an account that tells us that also tells us the conditions under which two sentences express the same proposition, thereby giving us a handle on propositional individuation. One wants to know what meanings ("semantic values," in the jargon of semantics) must be assigned to simple expressions (and syntactic structures) to determine propositions. Finally, one wants to know whether (the) propositions (an adequate semantics assigns to sentences) turn out to have structure (or at least individuation conditions) like that of the sentences that express them.

Begin with the last issue. A belief or assertion partitions possibilities into two classes, those in which it is true and the rest. One picture of propositions identifies them with such partitions. The view that propositions are sets of possible worlds is a version of this picture; so is the picture of propositions as sets of situations. Unlike most views which read the structure of a sentence onto the proposition it expresses, such views allow that difference in structure is not a bar to sentences saying the same thing: the claim that A&(B&C) is the claim that B&(C&A). And since a non-linguistic mental state may (in virtue of its relations to evidence and behavior) partition possibilities, this view makes ascription of belief to non-human animals relatively unproblematic.[16]

However, logically equivalent sentences are true in the same worlds. When A is a logical consequence of B, the worlds in which B is true are the worlds in which A and B are true. So it is impossible to know that B without knowing that A and B. The identification of propositions with sets of worlds does not solve the problem of hyperintensionality.

There have been some ingenious attempts to reconcile the idea that propositions are sets of worlds with the thought that one can believe what's said by

(a) 1 + 1 = 2

without believing what's said by

(b) 587/16 < 37.

Many depart from an observation of Stalnaker's that possible worlds play two roles in determining what is said.[17] Worlds provide the situations relative to which a sentence or thought token is true or false; they also help determine the semantic values of sentences and their parts. A toy example: Suppose we have fixed the reference of 'Hesperus' as the heavenly body that rises at position p, of 'Phosphorus' as the body that rises at position p′. In the actual world w, Venus rises at both p and p′, but in another world w′ different bodies may rise at p and p′, and in a third w″ (let's say) it is Mars that rises at both. In this case 'Hesperus is Phosphorus' determines a necessary truth at the actual world and at w″, a necessary falsehood at w′. It also, Stalnaker suggests, determines a "diagonal proposition," one true at world w* iff the sentence, interpreted in w*, is true at w*. This is a contingent proposition, true at w and w″, false at w′. Stalnaker suggests that in certain cases the content of a thought or assertion realized by a token of the sentence is the diagonal proposition. A general application of this proposal offers a way of simultaneously respecting the ideas that (for example): names rigidly designate their bearers; propositions are unstructured sets of worlds; although (a) and (b) interpreted according to "straightforward semantic rules" are true in the same possible worlds, they can express different thoughts.

Though ingenious, it is not clear that such views solve the underlying problem. For one thing, they are committed to the idea that anyone who knows that 1 and 1 is 2 knows that 587/16 < 37. True, on the above approach there is *an understanding* of 'anyone who knows that 1 and 1 is 2 knows that 587/16 < 37' on which it is false, at least given that sometimes *a knows that S* ascribes belief in the diagonal proposition determined by *S*. But one has to wonder why we should think that *any* construal of that sentence is true. Furthermore, such views tend to make the relation between the objects of belief and the bearers of truth and necessity obscure. It is, after all, necessary that 587/16 < 37. But then how can it be (sometimes truly said) that when John said that 587/16 isn't less than 37, he asserted something contingent – though, of course, it's not contingent that 587/16 < 37?[18]

Headway can be made if we assume that propositions have a structure which reflects the structure of the sentences which express them. If what's said by a sentence is individuated in terms of contributions made by the parts of the sentence, then logically equivalent sentences may, when they have different structures or their parts make different contributions, say different things.[19]

Given that propositional structure is derived from or at least reflected by sentence structure, it seems that the main problem an account of propositions has to answer is: What sorts of things are the constituents of propositions? A natural first answer is that they are the workaday semantic values of expressions: individuals, properties and relations, and things corresponding to connectives and logical operatives, if we select values as Russell would have us; possible-worlds intensions, if we select such values as does the possible-worlds semanticist.

This answer, however, seems to run into problems if we assume that what a sentence says is (roughly speaking) determined by "putting" semantic values contributed by expressions into the appropriate positions in a structure contributed by the sentence. There is no difference in the propositional structure contributed by 'Twain is dead' and 'Clemens is dead'; neither is there is any difference in the semantic values of the sentence's parts, on either of the above accounts of semantic values.[20] But it certainly seems that one can believe what one of the sentences says while not believing what the other does.

(Neo-)Russellianism and Fregeanism

Some say that this is wrong – you *can't* believe that Twain's dead without believing that Clemens is, for the proposition that Twain's dead just is the proposition that Clemens is. Such "direct reference" (sometimes called Russellian, Millian, or naïve) accounts of propositions have gained currency in good part as a result of arguments by Donnellan, Kripke, Putnam, and Kaplan against Fregean and descriptional accounts of the semantics of proper names and natural kind terms.

At the core of many of these arguments was the simple but compelling observation that, for example, what's said by 'Twain is dead' is, of necessity, true if and only if *Twain* is dead; what's said by 'Clemens is dead' is, of necessity, true if and only if *Clemens* is dead; and thus the truth-conditions of the claims that Twain is dead and that Clemens is dead are the same.[21] This makes it implausible, given the identity of what a sentence says and the object of the belief it expresses, that there is any (contingent) truth-conditional content which distinguishes the object of the belief that Twain's dead from that of the belief that Clemens is. But if there isn't this sort of difference between the objects of the beliefs, one might argue, it is not clear that there's any difference at all between them.[22]

Such arguments seem to give the lie to views of thought and its ascription like those held by Frege. Relevant to our present concerns are the following of Frege's views: Sentences and their significant parts have both *reference* and *sense*. An expression's reference is (roughly) what would nowadays be called its extension; its sense a "way of thinking" or "mode of presentation" of its reference. Frege's examples of senses of proper names are often given by associating definite descriptions with names, encouraging the widespread view that sense corresponds to some sort of descriptive conceptualization. In any case, it is the sense of an expression that is responsible for its having whatever reference it has. Senses are also the objects of propositional attitudes and the references of complement clauses. Thus, *a believes that S* is true just in case *a* bears the belief relation to the sense named by *that S*. The simple but compelling observation implies that the thought that Twain is dead cannot be identified with the thought that the F (for any F contingently true of just Twain) is dead. And this suggests that names are not synonymous with descriptions and do not have Fregean senses. But then it seems that the only contribution a name might make to what a sentence says is its referent. Given this and plausible principles of compositionality, the direct-reference view seems to follow.[23]

One might argue that if a speaker sincerely and understandingly dissents from a belief ascription, saying *I do not believe that S*, she must be correct about this; but this would not be so if the direct-reference view were correct. However, dissent does not imply disbelief. Suppose that A is watching B, who is across the street on the phone. A is also speaking to B on the phone, though A is not aware that the person seen is the person spoken to. Suppose

that A can truly say, pointing across the way, *I believe that she is happy*. Then B can say, to herself or through the phone, *The man watching me believes that I am happy*. If B can thus speak truly to A, A can speak truly, through the phone, to B, saying, *The man watching you believes that you are happy*. Since A can also say truly, *I am the man watching you*, a use of *I believe that you are happy* by A would also be true, even though A, we may suppose, would dissent.[24]

Direct-reference views, of course, conflict with strongly held intuitions of speakers. Speakers do not see the facts that Hesperus is Phosphorus, and that the ancients knew that Hesperus was Hesperus, as giving us any reason at all for thinking that the ancients knew that Hesperus was Phosphorus. Advocates of direct reference counter that these intuitions can be explained by distinguishing between what a sentence use literally says and what it might convey by non-semantic means. They suggest that (a) to have a propositional attitude is to be related to a Russellian proposition "under" a way of apprehending such; (b) while information about how a proposition is believed is typically conveyed by an APA, it is conveyed as a conversational implicature or via some other pragmatic, non-semantic, mechanism; (c) speakers' intuitions, while sensitive to the information a sentence use conveys, are often unable to distinguish between what is conveyed as a matter of truth-conditional content and what is conveyed pragmatically. (a) is made plausible by appeal to the above-mentioned attacks on Fregeanism. Suggestion (c) is independently plausible. And the distinction drawn in (b), between semantically and pragmatically conveyed information, is one which any comprehensive theory of language will have to draw. So unless we have an account that handles the data better than the direct-reference account, we should adopt it.[25]

I don't find this line of defense satisfactory, for it forces us to say that attempts to explain or predict behavior by ascribing propositional attitudes, if taken literally, *cannot* be successful. For it is the "way" a belief or desire is held that is relevant to its role in governing behavior, not (merely) its Russellian content. That Smith wants Twain dead, and that he believes that if he shoots, Twain will die, gives us not the slightest reason to think that Smith will shoot, given that the APAs are to be understood as the direct-reference theorist would have us understand them. For Smith might hold the desire under "Twain is dead," hold the belief under "if I shoot, then Clemens will die," and not accept "Twain is Clemens." Given the central role behavior explanation has in the practice of ascribing attitudes, it seems asking too much to ask us to relegate such explanations to the realm of pragmatic by-effects.[26]

A Fregean might simply concede to Kripke, Kaplan, Donnellan, and the rest that sense does not determine reference or the modally relevant properties of an expression.[27] She could still say that expressions have sense. She might say that sense is semantically irrelevant *except* in linguistic contexts sensitive to it: propositional-attitude constructions, those created by 'seek,' 'imagine,' and other such verbs, and a few others (e.g., 'means that'). Perhaps as an ecumenical gesture, the Fregean could allow that in such contexts expressions stand for something cobbled out of sense *and* reference. For example, perhaps 'Twain' in my use of 'Flo thinks that Twain is Clemens' would have as a reference the pair <Mark Twain, the sense I associate with 'Mark Twain'>. Thus, it might be said, we have the advantages of a Fregean account of attitude ascriptions without the drawbacks of an implausible account of reference.

Sense, however, is idiosyncratic, in so far as speakers may use a name to refer to an individual, but associate quite different senses with it. Suppose you have a belief you express with 'Frege was German.' I should be able to ascribe it to you by echoing you, saying, 'You think that Frege was German.' But if the complement clause names the thought *I* express

with 'Frege was German' it seems that the ascription won't be true, since you, having a different sense for 'Frege,' don't believe *my* thought that Frege was German (see Kripke, 1979; Richard, 1987).

One might suggest that in *a believes that S*, the complement names a sense which *a* associates with S. One problem with this is logical: It renders the argument *You think that Frege was German; Jo thinks whatever you do; so Jo thinks that Frege was German* invalid, since the complement now designates flaccidly. One might hold that a use of *that S* names the speaker's sense for S, and say that *a believes that S* is true iff *a* has a belief object that is *similar* to the one the complement names. But this again runs afoul of the fact that I can report the beliefs of others by echoing their words, as when the other expresses belief saying, 'Frege was German,' and I say, 'the other believes that Frege was German.' For I can do this even when the other thinks of Frege in ways quite unlike the way in which I think of him.[28]

Even if the Fregean account of attitude *ascription* is in error, a Fregean might say, still, attitudes *themselves* are relations to the sort of thing which Frege had in mind when he spoke of sense. But what are senses? A goodly number of Fregeans have suggested that at least in the case of proper names the notion of sense can be cashed out in terms of the notion of a *mental file or dossier*.[29] Roughly, the idea is that a way of thinking of something is a way of keeping track of, or a locus of information about, it. On this conception of sense co-referential names will have the same sense when their user directs information "tagged" with either to the same locus of information; thus, names a speaker knows to co-refer will typically have the same sense.

Whether this be Frege's notion or not, it is unclear that propositional attitudes can be individuated in terms of it. For one thing, assertions seem much more finely individuated: to say that Twain sleeps is not to say that Clemens does, even if speaker and audience have a single Twain/Clemens dossier. (For what one says does not shift simply because one's audience does; but one who understands what I say with 'Twain sleeps' and 'Clemens sleeps' may find only one claim informative.) Since there is pressure to identify the objects of beliefs and assertions (we can say what others think and think what others say), such considerations also count against individuating beliefs in terms of such a notion of sense.

Some contemporary Fregeans have argued that Frege's view of sense and reference is quite different from the view criticized by Kripke, Kaplan, and others. They say that the sense of a proper name is tied to what it presents in such a way that it is impossible for the sense to present anything other than that referent. As Gareth Evans explained the idea, my way of thinking of an object is characterized by describing how that mode of thought relates me to it; for you to think about something in the same way, the same description has to apply to you (save that references to me are replaced by references to you). Since the description refers to the object of thought, the way of thinking is one whose use (and thus existence) depends on its object.

A fanciful example: My sense for 'Frege' relates me to Frege because it is associated with (or is) information my teacher, Terry Parsons, conveyed to me, and which he gained by talking to Frege. It is impossible for such a description to be true of anyone unless Frege exists (and unless Terry Parsons exists, for that matter). So it's not possible for someone to think of something with my sense for 'Frege' unless these objects exist. Furthermore, if the description applies to you, it applies because you think of *Frege*, so my sense for 'Frege' can't pick out anything but him.[30]

It is commonly thought that a *virtue* of Frege's view, over that discussed in the last section, is that it can assign a thought to a sentence with an empty name, without committing itself

to the existence of a reference for the name. On Evans's account a name without reference is without sense, since there is no *x* such that the name represents a way of thinking of *x*. So sentences in which a referenceless name occurs have no sense, either. Whether this is ultimately objectionable depends in part on whether we should suppose (for example) that the same sort of explanation of Smith's behavior is to be given, when (1) Smith *sees* a cat, thinks "that's nice," and is moved to pet, and (2) Smith *hallucinates* a cat, thinks "that's nice," and is moved to pet.[31]

Attitudes, Utterances, and Sentences

If ascribing an attitude is not relating someone to a Russellian proposition or Fregean thought, what is it? Perhaps in ascribing attitudes we are talking about linguistic entities – sentences or utterances, for example. If so, we can in principle account for hyperintensionality. Even if 'I shout' and 'I yell' are synonymous, they are different sentences; to say that Mary is related to one needn't commit us to her being related to the other. Another attraction of such views is ontological: those suspicious of properties, possible worlds, or other "intensional entities" have hoped to make extensional sense of the attitudes by seeing them as relations to sentence tokens, utterances, or some other linguistic ersatz for propositions.

One well-known linguistic account of APAs is Davidson's paratactic one (Davidson, 1969). Davidson took the 'that' in 'says that' as a demonstrative, picking out the ensuing sentence-utterance. An indirect-speech report says that some utterance of its subject *samesays* the demonstrated utterance, with samesaying one or another relation of synonymy. Thus

[D] Derrida said that man is irrational

has the form and truth-conditions of

[Dl] Some utterance of Derrida's samesays that. Man is irrational.

where the demonstrative names the utterance of the sentence following. It is on Davidson's view no more the task of semantics to explicate the samesaying relation than to explicate relations picked out by other transitive verbs.

Lepore and Loewer observe that we can generalize this account to other propositional attitudes, by quantifying over states which stand to those attitudes as utterances stand to sayings (Lepore and Loewer, 1989). For example:

[S] Searle thinks that glass is transparent

has a logical form suggested by

[S1] There's a belief of Searle's which has-the-same-content-as that. Glass is transparent.

They suggest identifying belief states with neural ones.

The paratactic account is often thought to be committed to a manifestly false account of truth-conditions. If [D]'s truth-conditions are those of [Dl]'s, then [D] cannot be true unless

a particular English utterance exists, for [Dl] involves reference to such. But no utterance is such that Derrida could say that man is irrational only if it exists. Lepore and Loewer reply that u may samesay u′ even if u′ does not exist. This requires that utterances have their semantic properties essentially. If they are physical events, this may not be plausible. (For another response, see the end of this section.)

The paratactic account is fairly non-conservative as far as logical intuitions go. Burge points out that even arguments of the form *a believes that A. So, a believes that A* are not formally valid on Davidson's account, since the demonstratives they supposedly contain must vary in reference (Burge, 1986). The account has also been criticized on syntactic grounds, as it makes binding like that occurring in 'Every boy said that he is a fine fellow' mysterious.[32]

Perhaps the best-known account of ascriptions like [D] and [S] on which they are ascriptions of relations to sentence-types is Carnap's (1946). The account makes use of the notion of intensional isomorphism: roughly and slightly inaccurately, sentences are intensionally isomorphic provided they have the same syntactic structure and their simplest interpreted constituents have, pointwise, the same possible-worlds intension. Carnap's suggestion was (roughly) that [D] is true provided

> [D2] Derrida assertively uttered a sentence intensionally isomorphic to 'man is irrational'

[S] is true if, roughly, Searle is disposed to assent to some sentence intensionally isomorphic to 'glass is transparent.' These proposals are members of two large families of sententialist accounts of such ascriptions, other members being obtained either by replacing intensional isomorphism with some other relation between sentences (translation, for example), or by replacing the relation the subject of the attitude has towards the sentence most directly realizing her belief – replacing *is disposed to assent to* with *has a neural copy of*, for example.[33]

Church's objections to Carnap's proposal are taken by many to be decisive (Church, 1950). Church had two objections to Carnap. First of all, he complained that [D2] did not "convey the same information" as [D], since one gives the content of Derrida's assertion without revealing his words, while the other gives his words but not the content. Church reinforced the point by observing that literal translations of [D] and [D2] into German would clearly convey different things to a German speaker.

If we take Carnap's account as an attempt to spell out truth-conditions (thought of as sets of possible worlds or truth-supporting situations), this objection simply misses the point.[34] There is no reason to suppose that an illuminating account of the truth-conditions of (a use of) a sentence S and (the use of) S itself will convey the same information. For example, a correct account of the truth-conditions of 'some dogs bark' is given thus:

> Some of the things to which 'dog' applies are things to which 'bark' applies.

But this does not convey the same information as 'some dogs bark.' Indeed, an illuminating account (say, in terms of structure) of what proposition S expresses will typically fail to convey the same information as S.

Church's other objection to Carnap's account was in essence this. Example [D] and its German translation are not intensionally isomorphic, since they involve reference to different expressions (of English and German respectively). So 'Leo thinks that Derrida said that

man is irrational' and its German translation can be expected to vary in truth-conditions or even truth-value – Leo might be disposed to assent to an isomorph of [D] without being disposed to assent to an isomorph of its German translation. The objection assumes that a sentence and its translation can't diverge in truth-value, but surely this is false. 'He thinks Phil's a groundhog' and 'He thinks Phil's a woodchuck' may diverge in truth-value; but they are both translated by the same sentence in French, which has but a single word for the woodchuck.[35]

Church did point out a serious problem for *Carnap's* version of sententialism, but it is not clear that he uncovered a problem with *every* version of sententialism. For example, it is not clear why a sententialist who held

> *a believes that S* is true iff what *that S* names (the sentence S) translates a sentence "in a's belief box"

should be perturbed by Church's objection.

Some generic objections to sententialism should be mentioned in passing.[36] It's commonly alleged that sententialism is defeated by the fact that a sentence can mean different things in different languages. For all we know, there is a language L in which 'pigs fly' says that dogs bark. Thus 'pigs fly' translates (into L) a sentence ('dogs bark') which is in my belief box. Thus, on the account of the truth-conditions just displayed, it is true that I think that pigs fly, since the sentence 'pigs fly' translates a sentence realizing one of my beliefs. But I don't think that pigs fly.

There are natural ways of individuating sentences on which such objections have no force. We might identify words with sets of utterance tokens by speakers. English's 'pig' then turns out to be the set of all English 'pig' tokens. If sentences are structured collections of words, the objection fails, since L's words and English's words are surely different.

The second objection is this: If propositions are sentences, then that Smith thinks that fleas are disgusting entails that the sentence 'fleas are disgusting' exists. But Smith might have the belief without the sentence existing. A sententialist needn't accept the claim about entailment. Presumably there is a cross-world relation, *sentence S as used in w translates sentence T as used in w′*. S (in w) stands in this relation to T (in w′) in virtue of uses of S in w being similar in appropriate ways to those of T in w′. Such similarity does not require that S in fact exist in w′. (The contents of my bathtub don't have to exist in w′ in order for the underlying kind of my bathtub's actual contents to be the same as the underlying kind of some sample of liquid in w′.) A sententialist who endorsed a translational account of attitude ascription might simply say that *a believes that S* when used at a world w is true at world w′ just in case (1) there is at w′ a sentence T which realizes at w′ one of a's beliefs, and (2) S as used at w translates T as used at w′.

Semantic versus Psychological Sententialism

The view, that what *that S* provides in *a believes that S* is (individuated in terms of) a sentence, is a view about the semantics of belief ascription. It must be distinguished from the view that beliefs and other attitudes are realized in the mind or brain by states which are sentence-like in important ways, *psychological sententialism*, as it might be called.

On this view, the psychological states which realize propositional attitudes (a) have syntactic structure and parts with semantic properties like those of natural-language sentences; (b) have truth-conditions which can be assigned by something like a compositional induction on their structures. Views that we have a "language of thought," and must have such to be thinkers, are usually committed to this (and much more besides).

Psychological and semantic sententialism are independent doctrines. We could classify unstructured states with structured labels like sentences (semantic but not psychological sententialism). It is *a priori* possible that representing a state of affairs requires representing its constituents in a sentence-like way, though the point of attitude ascription is merely to report on how a person partitions possibilities (psychological but not semantic sententialism).

One line of argument for psychological sententialism claims that our best cognitive models of attitudes are ones on which they are computational states operating on sentence-like objects with properties (a) and (b).[37] Some have suggested that the "productivity" of the attitudes requires psychological sententialism: she who can entertain the thought that something bothered John, and can have the desire that Mary touch nothing, must also be able to entertain the thought that John touched something and be able to desire that Mary bother nothing. This is explained if the states are realized by a system of states with something like natural-language syntax and semantics; it is not so easy to see what other explanation there might be.

That the best models of attitudes are sentential is quite controversial.[38] Even without appeal to contentious psychological models, we should be leery of the above arguments. Attitudes involved in higher cognitive processes in humans are plausibly thought to be sentential for reasons given by these arguments. But one may wonder whether all belief (-like) states of monkeys (which are motivational, involve considerable discriminatory ability, but need not contribute much to something like an ability to reason) are well modeled by sententialist models.[39] Neither do such states obviously have the sort of "compositional complexity" alluded to above.

Opponents of psychological sententialism claim it is false to our conception of the attitudes, a conception of them as states with a certain functional role (usually a role captured by colloquial, or folk, psychology) and which have representational content. For it is perfectly possible to have representational content without being sententially structured; representational properties, for example, might be assimilated to one or another causal relational property. In my opinion, this argument ignores important aspects of our concepts of the attitudes. Attitudes *are* states with certain kinds of motivational roles and representational content; but an attitude is required to have a certain *kind* of representational content. An attitude represents a state of affairs – a congeries of individuals, properties, and relations – *by* representing its constituent individuals, properties, and relations.

Evidence that we think of attitudes as being realized by *structured* representations is given (a) by the naturalness with which we accede to the idea that, for example, having the belief that Bush lost to Clinton requires having a representation of Bush and one of Clinton; and (b) by the problems with accounts of attitude ascription which identify the semantic object of belief with an unstructured object like a set of possible worlds. We reject such accounts because we take it as obvious that states of affairs can be necessarily or logically equivalent without beliefs with those states as objects being identical. This means that the states which realize those beliefs must themselves be different. While it doesn't *follow* from this that belief states realize beliefs by being sentence-like representations of a state of affairs, such considerations, along with (a), make psychological sententialism plausible.[40]

Attitudes and Context

We have yet to find a satisfactory account of APAs. This section reviews some recent proposals. While they differ in particulars, they are united by the idea that attitude ascription involves some sort of *context sensitivity*, and that it is by appeal to this that we should explain hyperintensionality.

Kripke (1979) recounts the sad story of Pierre, who was raised in Paris but taken under dark circumstances to London. As a child, Pierre read (in French) of the city Londres, and accepted as true the French *Londres est jolie*, presumably expressing a belief therewith. He still accepts this sentence. Pierre learned English directly in London; it never occurred to him that he was in the city he called *Londres*. Finding his circumstances mean, he assents to 'London isn't pretty,' and presumably expresses a belief with it.

Kripke poses a puzzle about belief with the case. He observes that we seem committed to a disquotational principle about belief, something like *If a normal English speaker, on reflection, sincerely assents to 'p,' then he believes that p*. If this is true, its analogues in other natural languages are truths, too. We seem committed to the view that translations of truths express truths. The first principle, and the facts, commit us to saying that Pierre believes that London is not pretty. An analogue of the first principle, and the facts, commits us to the truth in French of *Pierre croit que Londres est jolie*. The translation principle, then, commits us to saying that Pierre believes that London is pretty. But, as Kripke observes, Pierre may be supposed to be fully rational, a man committed to consistency at all costs, and so on. So he presumably does not have contradictory beliefs.

Kripke seems to think that there must be a univocal, "context-free" answer to the question, Does Pierre or does he not believe that London is pretty? If we suppose that this is so, Kripke's case is indeed puzzling. For when we concentrate on Pierre's "French thoughts," it seems obvious that we should say that Pierre thinks that the city is pretty, and that he does not think that it is not pretty; when we consider his English thoughts, it is evident that he thinks that it is not pretty.

Perhaps Kripke's case is not so much a puzzle as a dramatic demonstration that ascriptions of propositional attitudes are contextually sensitive; perhaps the truth of 'Pierre thinks that London is pretty' depends in part on how things are with Pierre *and* in part on how things are in the context in which the sentence is used. A number of recent accounts of attitude ascriptions see claims about the attitudes as contextually sensitive. *Sentential-role* accounts take a use of

(P) Pierre believes that London is pretty

to be true just in case Pierre has a belief similar, in some contextually salient role, to that the user would normally voice with 'London is pretty.' *Translational* accounts see (P) as true, provided the complement clause (or what it names) provides an adequate translation of one of Pierre's thoughts; since what counts as an acceptable translation may vary from context to context, (P) may vary in truth across contexts. *Implicit reference* accounts see uses of sentences like (P) as involving a tacit reference by the speaker to a way of thinking, a mental particular (a particular representation or word in the language of thought), or a property of such. Since different speakers may refer to different ways of thinking or representations, uses of (P) may differ in truth.[41]

In Stephen Stich's version of the sentential-role story, belief states can resemble each other in terms of reference, functional role, and ideology, with ideological resemblance determined by the referential and functional similarities of the "networks" in which the states are embedded. Various similarity measures match such states along these dimensions; contextual differences in interest and emphasis select among these measures. Thus (P)'s truth-conditions shift, as do its users' attention and interest. Functional role and ideological setting might be seen as aspects of Fregean sense; Russellian propositions are a little like reifications of referential roles. So this view is a *bit* like the view that attitude ascriptions may relate one to either a Fregean sense or a Russellian proposition.

This approach makes our general success in explaining behavior by ascribing attitudes hard to explain. Presumably it is a belief's functional role that is most relevant in behavior explanation, and so similarity of such roles looms large in behavior explanation. But our behavior explanations are failures if they don't respect the referential properties of others' beliefs. Suppose we disagree about what's best for colds: you think pseudoepherine, I think oxymetazoline. Neither of us has much of a theory about his medication beyond thinking the stuff is the best for colds. 'I shall get some oxymetazoline' has for me the functional role that 'I shall get some pseudoepherine' has for you. Suppose that you are going to the store for pseudoepherine, Mary asks why you're leaving, and I say 'he wants to get some oxymetazoline.' If only similarity in functional role is relevant, this is true on Stich's account. But it's not true.

One response would say that referential similarity is always required in belief ascription, and, additionally, functional similarity is required in behavior explanation. But then I just *can't* explain your behavior to Mary, because nothing in my repertoire matches your motivation in reference and functional roles. But if I say 'he wants to get some pseudoepherine,' I explain your behavior quite well. If we say that *only* referential similarity is required in behavior explanation, the Russellian's problems with this topic recur.

Translational accounts of (P) take it as true when its complement provides a contextually adequate translation of one of Pierre's thoughts or something realizing them. The author's version of such an account begins by assuming that some (weak) version of psychological sententialism is correct.[42] Thus, each of Pierre's beliefs is realized by a sentence-like mental state whose parts, in context, determine a Russellian content. Call what results from pairing the parts of a "sentence" realizing a belief with their contextually determined referents a *thought*. Pretend for now that beliefs are realized by natural-language sentence tokens (which we think of as ordered sets). Then among Pierre's thoughts are

<<'est jolie,' the property of being pretty>, <'Londres,' London>>

and

<<'not,' the negation function>, < <'is pretty,' being pretty>, <'London,' London>>>.[43]

Call the pairings in thoughts of representations with what they represent *annotations*.

Thoughts are pairings of (the parts of) representations with what they represent. The complements of attitude ascriptions also determine such pairings. For example, the complement of (P) provides us with

<<'is pretty,' being pretty>, <'London,' London>>.

Call the pairings named by complement clauses *articulated propositions*, or a-propositions.[44] In ascribing an attitude, we offer the a-proposition our complement determines as a representation or translation of one of the believer's thoughts; the ascription is true provided that the proffered a-proposition is an acceptable translation of such a thought, according to currently prevailing standards.

What is translation, and how might the standards governing it shift across contexts? Translation preserves Russellian content.[45] So an a-proposition p translates a thought q only if p and q, when stripped of the words or representations within them, determine the same Russellian proposition.[46] What else is required in translation? This varies with interests, mutual knowledge, and conversational background. For instance, it may be common knowledge that some of Pierre's beliefs are realized *en français*, and are the focus of discussion. If so, one may expect a restriction on how to translate Pierre's thoughts along the lines of

> In discussing Pierre, 'London' can be used only to represent representations which Pierre voices with 'Londres.'[47]

With such a restriction in place, (P) will be true while

> (P1) Pierre believes that London is not pretty

will not. Other contexts provide restrictions which make (P1) true and (P) false.[48]

Call the sort of restriction just discussed – that in an ascription of attitude to x, an expression can only represent a representation of x's with a particular property P – a *restriction on translation*. Context contributes a collection of restrictions on translation. An ascription *A believes that S* is true in context c provided A has a thought q which can be translated (consistent with all c's restrictions) using the a-proposition that S.

This account attempts to preserve the virtues of direct-reference accounts – their compatibility with "the new theory of reference," their eschewal of "ways of thinking" as meanings or semantic values of expressions – with the idea that in ascribing an attitude we are *somehow* speaking of ways of thinking. It does this by holding that the words in a complement clause have two functions. One is to secure a Russellian referent: In (P), 'London' secures London. The other role is to stand as proxy for the representations of others: 'London' in a use of (P) is a proxy for a way of representing London.[49]

It is possible to get some of the effect of this account without requiring such yeoman service of the words in a belief ascription if we suppose that in using (P) we "tacitly refer" to Pierre's representations, but not via some expression in (P). Tacit reference is not unheard of; John Perry's example is the apparent reference to a location in uses of 'it's snowing.'[50] If (P) involves tacit reference to Pierre's representations, then the semantic value of the predicate

> (P3) believes that London is pretty

on a true use of (P) might be something like the property

> S_1: believing the Russellian proposition that London is pretty under the representation FR

where FR names Pierre's "French representation" of London's pulchritude. Since different uses of (P) may involve different references, simultaneous use of (P) may differ in truth-value.

So say Perry and Mark Crimmins. They hold that sometimes we refer to particular representations, sometimes to their properties. On some uses of (P), the semantic value of (P3) might turn out to be not S_1 but

> S_2: believing the Russellian proposition that London is pretty under a representation that includes a representation with property P*: being typically expressed by Pierre in French.

If there is reference to representations or their properties in belief ascription, then there will be reference to aspects of propositional structure and to how representations "fill" these. This is necessary to differentiate between believing that Hesperus rose before Phosphorus and believing that Phosphorus rose before Hesperus; it's only by saying which representation is responsible for filling which position (*role*, in Crimmins and Perry's parlance) that the difference between the two beliefs can be explained. A full-blown account of a use of (P) involving reference to P* has it making a claim something like this:

> There are representations r and k such that: Pierre believes the proposition that London is pretty under r, k is a part of r, and k, which has P*, is responsible for London filling role u

where role u is the "subject role" of the proposition that London is pretty.

Both the translational and the referential account, in effect, see attitudes as complex relations between an individual, a representation, and a representational content, with the latter being something like a Russellian proposition. Each sees attitude ascriptions as involving some hidden logical structure – quantification over ways of translating in the case of the translational account; reference to representations or properties thereof in the case of the referential account.

In my opinion, the implicit-reference account does not give an acceptable account of the logical properties of attitude ascriptions. Since the representations implicitly referred to may shift from premise to conclusion, the account must say that the argument 'Smith thinks snow is white, so Smith thinks snow is white' is not valid. Worse yet, my use of 'Smith thinks snow is white' may be false not because Smith fails to believe the proposition that snow is white, but because he does not believe it under the representation I refer to. Thus there are contexts in which 'Jones believes that snow is white' and 'Smith believes whatever Jones does' are true (taking the latter to be regimented 'For all p: if Jones believes p under some r, then Smith believes p under some r'), but 'Smith believes that snow is white' is false.

On the translational account, x satisfies *believes that London is large* just if 'London is large' adequately translates one of x's thoughts. Though the standards of adequacy in translation may vary from context to context, these standards are "built in" to the semantic value of the attitude-verb in a context. The verb's context-independent meaning is a rule which takes a collection of restrictions on translation, and returns a rule which pairs off individuals with the a-propositions which translate one of their thoughts. Because of the predicate's univocality in context, the arguments just discussed are valid on a translational account.

Such objections may not be decisive; one might hold that our intuitions about validity are no more infallible than those about truth-conditions.[51] I do believe that the translational or referential account gives the essentials of a correct account of attitude ascription.[52]

Alternatives to Relational Accounts

Alternatives to relational accounts of attitude ascription have recently appeared. We discuss the two most significant.

Relationism

In thought and talk we may represent an object multiple times, as when I think that Twain is Twain or when Mary assertively utters 'Twain isn't Clemens.' Several philosophers have suggested that such representational repetition is relevant to the individuation of meaning and content; Kit Fine has offered a sophisticated account of what he calls semantic relationism.[53]

According to Fine, uses of phrases stand in a relation of *coordination*, which is "the very strongest relation of synonymy" (Fine, 2007, p. 5). In a normal use of 'Twain is Twain' the uses of 'Twain' are coordinated with one another; in a normal use of 'Twain is Clemens' the uses of the terms are not coordinated. To understand use of the first sentence one must recognize the coordination; this signals a difference in meaning and proposition expressed. Similar things are true of uses of multiple sentences and multiple thoughts (even when the token uses or thoughts originate with different agents).

Fine thinks that the notion of coordination "can do much of the work of sense," even though (to take an example) on Fine's view there is no difference in proposition expressed by 'Twain is an author' and 'Clemens is an author' (Fine, 2007, p. 5). Fine is a referentialist and so in this case (where coordination doesn't enter the picture) sees no meaning-relevant difference. Fine posits an ambiguity (different "readings") of sentences like 'Ted said that Twain is an author.' On one reading, a use of the sentence is true iff Ted assertively uttered a sentence that expresses the proposition that Twain is an author *and* in so doing used a name of Twain that is coordinated with the ascriber's use of 'Twain.' Thus there is an understanding of 'Ted said that Twain is an author' on which its truth doesn't require the truth of 'Ted said that Clemens is an author.'

Whatever the virtues of the idea that content is relational, it's not clear that coordination can explain our practices of attitude ascription. Consider the following example: Eleanor Jane (=EJ) is Jane to her friends, Eleanor to others; her friends are well aware of this. You and I are her friends. I see that Bob and Ray, who know EJ but are not her friends, see her leave; I see that only Ray realized that it was EJ. (Of course, what Ray thinks is 'there goes Eleanor.') I say to you "Bob and Ray saw Jane leave, but only Ray knew/realized that it was Jane who left." This is a perfectly natural way to convey something along the lines of: Ray has some knowledge expressed by *a just left*, *a* a good answer for Ray to *Who's that?*; Bob has no such knowledge. An explanation of this in terms of coordination (as opposed, say, to one in terms of the translational account discussed above) presumably must invoke coordination between my use of 'Jane' and either some perceptual representation of Ray's of EJ or one of his tokens of 'Eleanor.' But the relation between my use of 'Jane' and Ray's representations doesn't seem to be anything at all like the relation Fine has in mind when he says, of coordination across people, that

> Perhaps a paradigm case is when I derive my use of a name from someone else…. there are many other cases … suppose … an object in common view and that we communicate about it by means of the pronoun 'it.' The various uses of the pronouns are coordinated and someone who did not know that they were being used to refer to the same object, even if he knew that each of them referred to that object, would have failed to understand what was said. (Fine, 2007, p. 86)

Someone might hear Ray say 'that's Eleanor leaving' and understand Ray, hear me say 'Ray knows that Jane left' and understand me while not knowing that Ray's 'Eleanor' co-refers with my 'Jane.' Someone who 'associates two routes' to EJ with the two terms might be like this.[54]

'Descriptivism'

Some separate the idea that belief is a relation to a proposition (or some other "belief object") from the idea that to ascribe belief is to say, of a person and a belief object, that the first is related to the second. Instead of using 'that Twain was an author' to *refer* to an object of belief, perhaps we use it to *describe* a belief object. Kent Bach, for example, proposes that a sentence of the form *x believes that S* is true iff that there is a proposition that x believes that has a property determined by *that S*. On this view, though the clausal complement in *x believes that S* expresses a proposition p, the role of this proposition is only to help determine a property P; the ascription is true only if some belief object of x possesses P.[55] What property the sentence ascribes varies with its use; phrases like 'believes that Twain is sad' are "semantically incomplete," and determine a property only when provided with contextual supplementation.

A motivation for this view is that it explains such things as why substitution of 'Clemens' for 'Twain' in the complement of an attitude ascription can change truth-value: uses of different complement clauses may determine different properties even if the clauses taken in isolation determine the same proposition.[56] It is not clear whether this view makes different predictions about the truth- or felicity conditions of uses of attitude ascriptions than does the translational view discussed at the end of the last section, particularly given that one way to specify a "translation manual" is in terms of properties of the representations that realize someone's attitudes.[57]

A variant descriptional view is given by Friederike Moltmann. Like Bach, Moltmann holds that in attitude ascription the clausal complement helps to pick out a property of the object of an attitude; unlike Bach, Moltmann takes the objects of the attitudes to be "products" of cognitive acts of judging, asserting, wishing, and so forth.[58]

Appendix: *De Dicto*, *De Re*, and *De Se*

It is common to distinguish between *de dicto* and *de re* attitude and attitude ascriptions. Those who do so may have any of a number of distinctions in mind. The most straightforward distinction is a syntactic one exhibited by

(la) John believes that Ned is tall.
(lb) John believes of Ned that he is tall.

and

(2a) John believes that Ned is tall.
(2b) There's some x such that x is Ned and John believes that x is tall.

In the b-sentences, the clause governed by 'believes' contains (elements like) variables which are bound from without. These are *de re* ascriptions. *De dicto* ascriptions, like the a-sentences, are ones in which the content sentence (the sentence which 'that' introduces) contains no variables bound from the outside.[59]

The terminology arises thus: *De dicto* ascriptions report that the believer has what a sentence says – a *dictum*, or proposition – as the object of an attitude. *De re* ascriptions relate the believer to a thing (or *res*) – Ned, in our examples – specified independently of how the believer conceptualizes it. Note that from (1b) and 'Ned is Ed' we can infer 'John believes of Ed that he is tall,' while (la) and the identity do not seem to imply 'John believes that Ed is tall.'

Say that when ascriptions are related as are

(3a) x believes that t is F
(3b) x believes of t that it is F

with 'it' in (3b) referring back to t, b is the *de re* ascription corresponding to a. Since arbitrary (positive) noun phrases can occur in the position of *t,* a *de dicto* ascription and the corresponding *de re* ascription are often independent. That you believe that some dogs have fleas neither implies nor is implied by the claim that you believe of some dogs that they have fleas. (You may have the first belief without the ability or willingness to identify or describe any particular thing as having fleas; you may identify Rex the dog as something beflead, but think him a skunk and that only skunks have fleas, thereby having the second belief but not the first.) Likewise, that you wish that the winner of the lottery be heavily taxed apparently neither implies nor is implied by the claim that you wish, with respect to the winner of the lottery, that he be heavily taxed.[60]

Limiting attention to cases in which *t* is replaced with something other than a quantifier – a demonstrative, indexical, or proper name – it is plausible that a *de dicto* ascription implies the corresponding *de re* ascription, for it is plausible that (for example) (lb) follows from (la).[61] Only the Millian holds that implication goes in the other direction.

Some accounts of attitudes make some attitudes dependent on objects external to the believer. A Russellian who takes Odile to be a part of the proposition that Odile said hello will usually hold that it's impossible to believe that Odile said hello if she does not exist.[62] The *de re* senses which some Fregeans introduce cannot exist unless what they (in fact) present exists. To have such a sense as a belief object is to have a belief which depends upon a particular object. A second use of the term *de re attitude* applies it to states which are object-dependent in this way – an attitude towards something (a Russellian proposition, *de re* sense, whatever) which is itself ontologically dependent upon an object which the belief is about.

This use of the terminology is different from the first. Sentence (la) reports a *de re* attitude on a Millian or Evansean view, since on these views the proposition that Ned is tall cannot exist unless Ned does. But sentence (la) is not a *de re* ascription.

A third use of the terminology arises as a result of the view that some belief involves an epistemically significant relation to an object. Russell held that if a proposition contained an object as a constituent, one couldn't believe the proposition unless one were *acquainted* with that object; acquaintance, in turn, could only be had with objects about which one, in some important sense, could not be deluded (see Russell, 1911; 1912). Even though no one now accepts quite so stringent a requirement, it is commonly thought that object-dependent propositions can be believed only by those with some fairly significant epistemic contact with their constituents.[63] Many writers use '*de re* belief' in such a way that *de re* belief is a kind of belief whose possession requires having one or another epistemically interesting rapport with the object or objects the belief is about.

Does having an object-dependent belief (*de re* in the second sense) require having intimate epistemic contact with the relevant objects (*de re* in the third)? I would say not, as anyone who understands a sentence with a proper name of me may have object-dependent beliefs about me, but such a person need have no particularly interesting epistemic relation to me. An objection runs thus: On your view, someone without such contact with an object x – say, someone one who can (only) describe x – can come to have object-dependent beliefs about it just by "christening" x with a name stipulated to always "introduce" x into the belief it expresses. This implies that someone with only descriptive knowledge of x can be in exactly the same epistemic relation to x as someone who has object-dependent knowledge of x. But this is absurd, since someone who has object-dependent knowledge knows something that someone with only descriptive knowledge doesn't.

This argument succeeds only if it is allowed that introducing a name for an object does not change epistemic relations to the object. But introducing a name (when one previously had none) does create a new epistemic link with the nominata: after the christening, one has a means to *refer* to the object, and express propositions in which the object occurs, while before one did not. Of course, the sort of link a christening opens to the object can't be exploited to gain "interesting extensions" of one's knowledge in the way that other epistemic links (perception, introspection, relations established via third-party testimony) can be. But that doesn't mean that the link doesn't exist, as the argument must assume in order to go through.

There is apparently a difference between what is typically reported by

(4a) John believes that he himself is tall

and

(4b) John believes that John [or: that man] is tall,

even if 'John' (or 'that man') refers to John himself. If John doesn't realize that he is John (or that man) – because of amnesia or some other circumstance – the a- and b-sentences might diverge in truth-value. For if John is suffering from amnesia, he can think (to himself) *I am tall*, while thinking *But John (that man) is not tall*. Assuming that inferences like that from (1a) to (1b) are valid, this argument also shows that what's reported by (4a) isn't (merely) a *de re* belief, either. The difference is even more pronounced with attitude verbs that take infinitival complements. Witness the difference between

(5a) Hazel expected to meet me for lunch

and

(5b) Hazel expected that Hazel would meet me for lunch.

The term *attitude de se* is often used to refer to the sort of belief about oneself typically reported in English using the 'she herself' locution or by attitude verbs with an infinitival complement.[64]

What, exactly, is (4a) telling us about John? According to Frege, "everyone is presented to himself in a particular and primitive way, in which he is presented to no-one else" (Frege, 1977). In saying this, Frege had in mind a "distinctively first-person" way of thinking of

oneself, which typically accompanies one's 'I'-thoughts.[65] One view is that a *de se* ascription to x ascribes to x a thought which involves x's private mode of self-thought.[66]

One way to avoid private thought-objects in an account of thought *de se* treats 'he himself' (and analogous uses of simpler pronouns like 'he') as functioning somewhat as predicate abstractors, so that the logical form of (4a) is suggested by

(4c) John believes λx(x is tall)

One then says that *de se* belief involves a distinctive way of ascribing a property (self-ascribing it). All believers can self-ascribe properties, and all properties are (in principle) open to self-ascription. But of course only you can *self*-ascribe being tall to yourself.

As noted above, David Lewis takes this line, and holds that *all* attitudes are relations to properties; what appears to be "purely propositional belief" (such as the belief that 2 + 5 = 4) is self-ascription of "propositional properties" (being such that 2 + 5 = 4). Lewis argues that this provides a superior account of belief, using examples like that of the propositionally omniscient gods, Gary and Ed. Those working in a possible-worlds framework often adopt a variant of Lewis's account, on which the objects of attitudes are sets of "centered worlds," where a centered world is a triple <w, u, t>, w a possible world, u an object in w, and t a time; what self-ascribing (say) the property of being sad comes to, if one adopts this framework, can be identified with having {<w,u,t> | at t in w u is sad} as an object of belief.[67]

Lewis's and kindred examples raise a number of questions. Are *de se* attitudes a distinctive sort of attitude that cannot be identified with or somehow reduced to *de re* and/or *de dicto* attitudes? Are there forms of attitude ascription (or distinctive understandings of some such forms) in natural language that are distinctively *de se* and which convey information that is not conveyed by *de re* and/or *de dicto* attitude ascriptions? If the answer to either of these is affirmative, does an account of *de se* attitudes or *de se* ascriptions require us to introduce a new sort of content? It has been argued that examples like Lewis's give us no good reason to think that the answer to the first two questions is affirmative.[68] But even assuming that we think that there are distinctively *de se* attitudes and ascriptions, it is not clear that that is reason to think that content is not propositional in some central sense. Fregeans argue that they can accommodate the *de se* in terms of special modes of presentation.[69] A translational account of attitude ascriptions can allow that for any sentence S which is free of the 'he himself' locution (or cognates), if S is true, then 'Ed knows that S' is true in the two-gods story. This seems to get at what Lewis has in mind in describing the case as one in which Ed is propositionally omniscient.[70] And it is consistent with the falsity of 'Ed knows that he himself is Ed,' if 'he himself' is so restricted that in ascriptions to Ed it can only represent 'I' or other "first-person" modes of reference. Other views of the nature of content can accommodate Ed and Gary's ignorance in other ways.[71]

Notes

1 The caveat 'normally' allows us to hold that one might believe, say, that Zeus was a god, though the belief would be without truth-conditions due to a reference failure of 'Zeus' or a corresponding part of the representation realizing the belief. It is unclear that we can avoid such a rider, since there need not be only finitely many ways in which a representation might fail to have truth-conditions. The parenthetical 'or the like' allows, for instance, that attitudes like desire and wishing may have objects with satisfaction, not truth-, conditions.

2 Chomsky (1965; 1986) holds that we know the grammar of our language, though the knowledge is not conscious or inferentially integrated with conscious knowledge. Stich (1978) criticizes him on these points. Fodor (1975) exuberantly postulates propositional attitudes without regard to conscious access or cognitive integration; Searle (1990) holds that representation without the possibility of conscious access is impossible.

3 See, for instance, Higginbotham (1983), which presents an alternative to accounts in Barwise (1981) and Barwise and Perry (1983). Barwise and Perry take seeing a lion dance to be a relation between a "scene" and a perceiver, but (in 1983) assimilate the objects of the attitudes to the objects of perception. Neale (1988) discusses Higginbotham's account.

Martin (1992) argues that visual perception may involve representation without (so to speak) the kind of conceptualization required for a propositional attitude, as when I see an object without immediately registering its presence or properties.

4 Attitude verbs presumably pick out relations between individuals, propositions, and times; strictly, attitudes themselves are such relations. For simplicity, tense and time are ignored throughout.

5 There are important differences between assigning semantic properties to expression types in a context and assigning them to expression tokens or their utterance. Most of the sequel slurs this distinction. And I generally suppress reference to the fact that expressions are sensibly assigned semantic values only relative to a context of use.

Frege, Russell, and many of their heirs seem to have assumed that sentence uses express at most one proposition. Many contemporary theorists have abandoned this idea (see, for example, Cappelen and Lepore, 2005), though those of a Gricean bent will say that one of these – what is strictly and literally said – is distinguished via its relation to a sentence's meaning.

Many now say that typically a sentence's meaning does not determine a proposition-cum-object-that-determines-truth-conditions, but something – a propositional schema, or propositional skeleton – that requires supplementation before truth-conditions can be determined. (A standard example is a sentence like 'My book is interesting': the possessive is not ambiguous, but it can express many relations (here, for example, authorship and possession).) For discussion see Cappelen and Lepore (2005), Bach (1994), and Soames (2002). Those who hold such views will hold at best a qualified version of what is below called 'the relational account.' It would take us too far afield here to discuss this variant view in detail.

6 Italicization is used as a device for talking about expressions. (Precisely: It functions as a method of quasi-quotation, in Quine's sense.) Single quotes are used to mention expressions.

7 This line of thought can be found in Salmon (1986) and Soames (1987).

8 See Higginbotham (1992), Soames (1992) for discussion.

9 Davidson (1967). Some identify the task of a theory of meaning with giving an account of truth-conditions, which they in turn identify with assigning sets of worlds or situations to the sentences, which are in turn identified with what sentences say. See, for example, Montague (1974).

10 Hyperintensionality is distinct from what Quine calls *opacity*. As Quine defines the notion, a linguistic context $e(\ldots)$ is opaque provided there are singular terms t and t' such that $t = t'$ is true, but $e(t)$ and $e(t')$ have different extensions. The opacity of 'believes that… won' follows from the assumptions that definite descriptions are terms, and that 'the lottery winner = the man in the corner,' 'Jo thinks that the lottery winner won,' and 'Jo doesn't think that the man in the corner won' are all true. That 'believes that… won' is hyperintensional does not follow from this. Conversely, the hyperintensionality of 'believes that this is…' can be demonstrated by means of a suitable pair of necessarily equivalent predicates, but its opacity can't.

11 Precisely: Transformations which are valid within the scope of modal operators fail within the scope of attitude verbs.

12 See Mates (1950) and Burge (1978); compare Yagisawa (1984).

13 For Soames's view see Soames (2010b), for King's, King (1997). These and related views are critically discussed in a collection edited by Hunter and Rattan (2013).

14 Of course those who say this need to explain how it can be that, for example, it's necessary that the proposition that 7 is prime is true – and thus would be true even if there were no things to represent. For one account of this see Richard (2013a; 2013c).

15 The question of whether propositions can change truth-value over time is the subject of Brogaard (2012); some reasons to doubt Brogaard's view can be found in the first part of Richard (2015). Discussion of the sort of relativism suggested by (c) can be found in Capellen and Hawthorne (2010), MacFarlane (2014), and Richard (2011).

16 On propositions as sets of worlds, see Stalnaker (1984; 1987; 1999), Powers (1978), Richard (1990), and Cresswell (1985). Soames (1984; 1987) objects to situation semantics' account of propositions; some of his points are anticipated in Richard (1983). Stalnaker (1984) and (1999) give several arguments for the view that the objects of belief should not be individuated more finely than sets of worlds. Some of these arguments are critically discussed in Richard (2013b). An alternative account identifies propositions with collections of possible and impossible worlds (or situations). See Soames (1987) for criticism and Edelberg (1994) for a defense.

17 See the essays in Stalnaker (1999). An elaborate defense of this sort of idea is found in the work of David Chalmers; see, for example, Chalmers (2002).

18 See Richard (2013b, ch. 9), where this sort of point is made about a view Graeme Forbes once held. Forbes presses this kind of objection against views of David Chalmers in Forbes (2011).

19 Lewis (1972) and Cresswell (1985) are possible-worlds semanticists using structured propositions. Salmon (1986) and Soames (1987; 1989) adopt a Russellian view on which the structure of the proposition that S generally recapitulates that of S itself. Of course, many linguistic views of propositions identify propositions with structured (linguistic or quasi-linguistic) items; among the many examples are Segal (1989), Richard (1990), and Larson and Ludlow (1993).

20 Given the almost universally accepted claim that names rigidly designate their bearers; for discussion see Chapter 36, NAMES AND RIGID DESIGNATION, and Chapter 35, REFERENCE AND NECESSITY.

21 A little more precisely, the arguments made it plausible that such terms are rigid designators of their bearers, and thus the modal properties of the propositions expressed by *A(t)* and *A(t′)* are identical if t and t′ name the same thing and the dotted position in *A(…)* does not occur in the scope of a device of quotation or attitude verb. See Kripke (1972), Kaplan (1977), Donnellan (1972; 1974), and Putnam (1975a).

22 In this argument, 'object of belief' should be understood as shorthand for 'what's named by the complement clause in a belief ascription.'

23 But see Crimmins (1992) for discussion of this line of argument.

24 This argument is in Richard (1983). Soames (1987) suggests that it generalizes to show that co-referential proper names are intersubstitutable in APAs. Forbes (1987), Richard (1990), and Crimmins (1992) offer various suggestions as to why the argument does not establish a direct-reference view.

25 Such views owe a great deal to Grice (1990) and Kripke (1977). Versions are in Richard (1983; 1987); Soames (1984; 1987); Salmon (1986); and Berg (1988).

26 Richard (1987) tries to square folk psychology with the direct-reference view. Richard (1990) and Crimmins (1992) argue that the view's inability to account for the literal truth of folk psychology is a serious flaw.

27 Graeme Forbes (1989) has suggested that sense determines actual reference, and that contexts like necessarily … are extensional, the extension of a sentence being a state of affairs. He then adopts an account of sense like that of Evans discussed below. Richard (1993a) and Crimmins (1993) discuss Forbes's account.

28 See Richard (1988; 1990). One hears the response that (a) I cannot understand the other unless I know that she uses 'Frege' to refer to Frege, but (b) I cannot know this unless there is some similarity in the way the other and I think of Frege. But (b) is wrong. For example, I might know on the basis of third-party testimony that you use 'Frege' for Frege; or I might know that you are an

American and in situation F, and that Americans in situation F use 'Frege' for Frege. It is also unclear why I *must* understand the other in order to correctly ascribe a belief to her.

29 See Evans (1982) and Forbes (1989). The idea arguably originates in Kaplan (1969).

30 See Evans (1982) and McDowell (1984). Such views, of course, seem heir to the problems about the idiosyncrasy of sense mentioned above.

31 For discussion, see Evans (1982), Segal (1989), and Recanati (2013).

32 See Higginbotham (1986) and Hand (1991). Hand also observes that for many cases involving negative polarity and infinitives, positing (the utterance of) a discrete sentence with the content of the ascribed attitude is implausible.

33 Intensional isomorphism (and relations like translation) are not dyadic but actually quadratic relations between two sentences and two languages.

34 Admittedly, Carnap himself did not seem to take it in this way.

35 See Richard (1990). A different response is given in Leeds (1979).

36 See Schiffer (1987) for an inventory. Richard (1995a) attempts to meet serious objections to sententialism in a way that would be acceptable to those with broadly nominalistic inclinations.

37 Perhaps the most influential arguments are in Fodor (1975; 1987); those mentioned in the text are suggested by him. A sampling of the literature is in Block (1981) and Rosenthal (1991).

38 McLaughlin (1993) gives an introduction to current debate.

39 A summary of some relevant observation and experiment is Cheney and Seyfarth (1990).

40 For discussion see Stalnaker (1984), Dennett (1987), Richard (1990), and Crimmins (1992).

41 Stich (1983) and Boer and Lycan (1986) present sentential role accounts. The latter is discussed in Richard (1990). Richard (1989; 1990; 1993a; 1993b) develops a translation account. Grandy (1986) offers an earlier translational account. Schiffer (1979) suggests an implicit reference account, in which tacit reference is (apparently) made to intersubjectively accessible ways of thinking of objects and properties. Perry and Crimmins (1989) and Crimmins (1992) develop the version discussed below. Richard (2013b, ch. 14) contains a lengthy discussion of Kripke's puzzle.

42 See Richard (1989; 1990; 1993b).

43 Here the quotation names should be thought of as names of particular token representations of Pierre's which are instances of the quoted types.

44 Called *Russellian Annotated Matrices* in Richard 1990. Note that in describing a-propositions, quote names name expression types.

45 It would be possible to deny this and still preserve something of the spirit of the account suggested here. One might say that the normal or default mode of translation is one which preserves Russellian content. One would go on to suggest that just as the intentions of speakers might require a use of 'London is pretty' in 'Pierre thinks London is pretty' to represent one of Pierre's 'French thoughts,' so might these require a use of 'this is a very pleasant place' to represent a thought expressed by 'hier ist es sehr gemütlich,' though the Russellian content of 'gemütlich' and 'pleasant' in their respective languages are not identical.

Two developments suggest themselves. The less radical simply allows that in context, correlations (the functions which map parts of a-propositions to parts of thoughts) may occasionally map (e, x) to (f, y) when x is not y. More radically, the atomistic account of translation used in the text would be replaced by a more holistic one, in which correlations are replaced by functions which map whole a-propositions to whole thoughts.

46 Slightly more precisely: Translation requires that there be a *correlation* function f which maps the annotations in a-propositions to annotations in thoughts such that q can be obtained from p by replacing p's parts with their image under f. The text suppresses complications arising with iteration of attitude verbs. See Richard (1990).

47 Strictly speaking, a restriction on translation tells one that an annotation (such as the pair <'London,' London> can only be used to represent certain kinds of annotations (e.g., those of the form <'Londres,' London>. For details see Richard (1990; 2013b).

48 The restriction in the text is really on the use of the annotation <'London,' London>. A more detailed account of the nature of contextual restrictions is in Richard (1993c). There it is proposed that context associates a (possibly null) "theory" about how a representation functions in thought with each pair of an individual u and expression e; the contextual restriction on a use of e in an ascription of attitude to u is that e can be used to represent only a representation r of u's such that the theory associated with u and e is substantially correct when taken as a theory about r.

49 Soames (2002) gives useful criticisms of the view just sketched. A response to Soames is given in Richard (2013b, ch. 13).

50 See Perry (1986). Perry and Crimmins (1989), and Crimmins (1992) give the account sketched below.

51 For critical discussion of translational and implicit reference accounts, see Crimmins (1992), Richard (1993b; 1995b), Saul (1993), Schiffer (1992), and Soames (1995).

52 Later work by Crimmins (1998) pursues a very different account of attitude ascription, on which it involves pretense. For example, when one says that someone believes that Hesperus is not Phosphorus, one pretends that Hesperus and Phosphorus are different things, so that thinking about Hesperus is not the same thing as thinking about Phosphorus. Crimmins's account exploits ideas of Kendall Walton about pretense and utterance truth in order to assign intuitively correct truth-conditions to attitude ascriptions. It would add too many words to an already too lengthy chapter to discuss Crimmins's imaginative account; Richard (2013b, ch.10) gives a critical discussion.

53 Fine (2007). There is discussion and elaboration of Fine's view in Soames (2010a) and Fine (2010).

The idea that (for example) there is a difference in content between sentences of the forms *S(…t…t…)* and *S(…t…t′)* even when t and t′ are synonymous is not new. Putnam (1954) (reprinted in Salmon and Soames, 1988) suggests something like this; chs 2–4 of Richard (2013b) attempt to develop a view that is in some ways kindred to Fine's.

54 Of course such a person wouldn't know that Ray's utterance is what grounds the truth of my utterance, but that seems neither here nor there so far as Fine's elaboration of the notion of coordination is concerned.

55 Bach (1997). As I understand Bach, P is possessed by q only if q entails p.

56 Bach does not give an account of the logical syntax of attitude ascriptions, so my exposition may make presuppositions not wholly faithful to his intent.

57 Because of space considerations, I do not here take up Michael Devitt's (1995) more elaborate descriptivist view. Unlike Bach, Devitt gives a catalog of the sorts of properties that he takes to be (quasi-)conventionally associated with clausal complements; he also (unlike Bach) attempts to give an explanation of how it could be that the meanings of attitude ascriptions are consistent with the idea that folk-psychological belief/desire generalizations are in some sense lawful. For a critical examination of Devitt's view, see the exchange between Devitt and me in Richard (1997) and Devitt (1997).

58 Moltmann (2014). Like Devitt, Moltmann essays a catalog of the sorts of properties that are ascribed by use of clausal complements. Moltmann adduces some interesting syntactic evidence for the idea that attitude ascriptions are ascriptions of relations to "cognitive products."

59 Slightly more precisely, *de re* ascriptions contain elements (pronouns or variables) which are within the scope of the complementizer for the attitude verb, and which are either bound to a noun phrase outside the complementizer's scope, or have an antecedent outside the complementizer from which they acquire their reference. (This definition may not capture quite the class of ascriptions each author who uses the terminology calls *de re*.)

60 This has been denied by various latitudinarian accounts of *de re* attitudes; see Chisholm (1976) for an example. See Kaplan (1969) and Sleigh (1967) for discussion.

61 I take definite descriptions to be quantifiers.

62 This assumes that it's impossible to have the belief unless the proposition exists, and that it's impossible for a proposition to exist unless its parts do.

63 Much of the discussion of Kripke's examples of contingent *a priori* knowledge seems to be driven by the assumption that object-dependent knowledge requires epistemic rapport. See, for example, Donnellan (1979) and Forbes (1989). See also Kaplan (1969; 1977; 1978; 1989) and Richard (1993a). Jeshion (2010) contains recent work on object-dependent knowledge.

64 Castenada (1966; 1967) first brought out the importance of the divergence between the a- and b- sentences. It has been extensively discussed in the literature; early examples are Perry (1979), Lewis (1979), and Chisholm (1981). Boer and Lycan (1980) deny that there is a difference in the truth-conditions of (4a) and (4b); Boer and Lycan (1986) retract this.

A standard assumption in the linguistics literature is that in sentences like (5a) the complement contains an unvoiced phrase (usually called PRO) in subject position that is controlled by the ascription's subject term. (To a first approximation, control is a relation in which the controlled element inherits its semantic properties from the controller.) It has been suggested that when an attitude verb has its infinitival complement with a PRO subject controlled by the verb's subject, the only interpretation possible is *de se*. For discussion see Chierchia (1990) and Pearson (2012).

65 Why, one might ask, must this way of thinking be private? According to Frege, a way of thinking presents no more than one thing. Presumably Frege assumes that the way of thinking in question would 'seem first-person' to whomever used it – so that if I were to use the mode of thought you use with T to think of x, I would have to do the sort of thing I do when I think to myself *I am tall*. This all seems to imply that if I could use your private mode of self-thought, I could think that you are tall by thinking *I am tall*, which seems absurd.

66 Evans (1981), Peacocke (1981), McGinn (1983), and Forbes (1987) present Fregean accounts of *de se* thought.

67 There are epicycles on this that one might invoke if one works in a Hintikka-style framework. For discussion see Pearson (2012).

68 A recent example is Cappelen and Dever (2014).

69 Evans makes suggestions about how one might do this in Evans (1982); a variant of Evans's approach is discussed in Stanley (2011).

70 This ignores presumable (but irrelevant) limitations of expressive capacity in English.

71 Examples are the discussion in Stalnaker (2008, ch. 3) and that in Perry (1979). Of the later, Lewis (1979) says that he is sure it "works as well" as his own, but that it is "more complicated."

References

Almog, J., J. Perry, and H. Wettstein, eds. 1989. *Themes from Kaplan*. Oxford: Oxford University Press.

Bach, K. 1994. "Conversational impliciture" *Mind and Language*, 9(2): 124–162.

Bach, K. 1997. "Do belief reports report beliefs?" *Pacific Philosophical Quarterly*, 78(3): 215–241.

Barwise, J. 1981. "Scenes and other situations." *Journal of Philosophy*, 78(7): 369–397.

Barwise, J., and J. Perry. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.

Berg, J. 1988. "The pragmatics of substitutivity." *Linguistics and Philosophy*, 11(3): 355–370.

Block, N. 1981. *Readings in the Philosophy of Psychology*, vols 1 and 2. Cambridge, MA: Harvard University Press.

Boer, S., and W. Lycan. 1980. "Who, me?" *Philosophical Review*, 89(3): 427–466.

Boer, S., and W. Lycan. 1986. *Knowing Who*. Cambridge, MA: MIT Press.

Brogaard, B. 2012. *Transient Truths*. New York: Oxford University Press.

Burge, T. 1978. "Belief and synonymy." *Journal of Philosophy*, 75(3): 119–138.

Burge, T. 1986. "On Davidson's 'Saying that.'" In *Truth and Interpretation*, edited by E. LePore, pp. 190–208. Oxford: Blackwell.

Cappelen, H., and J. Dever. 2014. *The Inessential Indexical*. Oxford: Oxford University Press.

Cappelen, H., and J. Hawthorne. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cappelen, H., and E. Lepore. 2005. *Insensitive Semantics*. Oxford: Blackwell.

Carnap, R. 1946. *Meaning and Necessity*. Chicago: University of Chicago Press.
Castenada, H.-N. 1966. "'He': a study in the logic of self-consciousness." *Ratio*, 8: 130–157.
Castenada, H.-N. 1967. "Indicators and quasi-indicators." *American Philosophical Quarterly*, 4: 85–100.
Chalmers, D. 2002. "On sense and intension." In *Philosophical Perspectives*, vol. 16, edited by J. Tomberlin. Atascadero, CA: Ridgeview.
Cheney, D., and R. Seyfarth. 1990. *How Monkeys See the World*. Chicago: University of Chicago Press.
Chierchia, G. 1990. "Anaphora and attitudes de se." In *Semantics and Contextual Expression*, edited by R. Barsch, J. van Benthem, and P. van Emde Boas, pp. 1–32. Dordrecht, Netherlands: Foris.
Chisholm, R. 1976. *Person and Object*. La Salle, IL: Open Court.
Chisholm, R. 1981. *The First Person*. Minneapolis: University of Minnesota Press.
Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Prager.
Church, A. 1950. "On Carnap's analysis of statements of assertion and belief." Reprinted in Linsky, 1971, pp. 168–170.
Cresswell, M. 1985. *Structured Meanings: The Semantics of Propositional Attitudes*. Cambridge, MA: MIT Press.
Crimmins, M. 1992. *Talk about Beliefs*. Cambridge, MA: MIT Press.
Crimmins, M. 1993. "Forbes' so-labelled neo-Fregeanism." *Philosophical Studies*, 69(2): 265–279.
Crimmins, M. 1998. "Hesperus and Phosphorus: sense, pretense, and reference." *The Philosophical Review*, 107(1): 1–48.
Davidson, D. 1967. "Truth and meaning." In Davidson, 1984, pp. 17–36.
Davidson, D. 1969. "On saying that." In Davidson, 1984, pp. 93–108.
Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.
Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
Devitt, M. 1995. *Coming to Our Senses*. Cambridge: Cambridge University Press.
Devitt, M. 1997. "Meanings and psychology: a response to Mark Richard." *Noûs*, 31(1): 115–131.
Donnellan, K. 1972. "Proper names and identifying descriptions." In *Semantics for Natural Language*, edited by D. Davidson and G. Harman, pp. 356–379. Dordrecht, Netherlands: Reidel.
Donnellan, K. 1974. "Speaking of nothing." *Philosophical Review*, 83(1): 3–31.
Donnellan, K. 1979. "The contingent *a priori* and rigid designators." In *Contemporary Perspectives in the Philosophy of Language*, edited by P. French, T. Uehling, and H. Wettstein, pp. 45–60. Minneapolis: University of Minnesota Press.
Edelberg, W. 1994. "Propositions, circumstances, and objects." *Journal of Philosophical Logic*, 23(1): 1–34.
Evans, G. 1981. "Understanding demonstratives." In *Meaning and Understanding*, edited by H. Parret and J. Bouveresse, pp. 280–303. Berlin: Walter de Gruyter.
Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.
Fine, K. 2007. *Semantic Relationism*. Oxford: Wiley-Blackwell.
Fine, K. 2010. "Comments on Scott Soames's 'Coordination Problems.'" *Philosophy and Phenomenological Research*, 81(2): 475–484.
Fodor, J. 1975. *The Language of Thought*. New York: Thomas Y. Crowell.
Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
Forbes, G. 1987. "Indexicals and intensionality: a Fregean perspective." *Philosophical Review*, 96(1): 3–31.
Forbes, G. 1989. *Languages of Possibility*. Oxford: Blackwell.
Forbes, G. 2011. "The problem of factives for sense theories." *Analysis*, 71(4): 654–662.
Frege, G. 1977. "The thought." In *Logical Investigations*, translated by P. Geach and R. Stoothoff. New Haven: Yale University Press.
Grandy, R. 1986. "Some misconceptions about belief." In *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, edited by R. Grandy and R. Warner, pp. 317–331. Oxford: Oxford University Press.

Grice, P. 1990. *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.
Hand, M. 1991. "On saying that again." *Linguistics and Philosophy*, 14(4): 349–365.
Higginbotham, J. 1983. "The logic of perceptual reports: an extensional alternative to situation semantics." *Journal of Philosophy*, 80(2): 100–127.
Higginbotham, J. 1986. "Linguistic theory and Davidson's program in semantics." In *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, edited by E. LePore, pp. 29–48. Oxford: Blackwell.
Higginbotham, J. 1992. "Truth and understanding." *Philosophical Studies*, 65: 3–21.
Hunter, D., and G. Rattan, eds. 2013. "Essays on the nature of propositions." Special issue of *Canadian Journal of Philosophy*, 43: 5–6.
Jeshion, R., ed. 2010. *New Essays on Singular Thought*. Oxford: Oxford University Press.
Kaplan. D. 1969. "Quantifying in." In *Words and Objections*, edited by D. Davidson and G. Harman, pp. 206–242. Dordrecht, Netherlands: Reidel. Reprinted in Linsky, 1971, pp. 112–144.
Kaplan. D. 1977. *Demonstratives. Draft* #2. Dittograph. Reprinted in Almog, Perry, and Wettstein, 1989.
Kaplan. D. 1978. "Dthat." In *Syntax and Semantics*, vol. 9, edited by P. Cole. New York: Academic Press.
Kaplan. D. 1989. "Afterthoughts." In Almog, Perry, and Wettstein, 1989.
King, J. 1997. *The Nature and Structure of Content*. Oxford: Oxford University Press.
Kripke. S. 1972. *Naming and Necessity.* In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 253–355. Dordrecht, Netherlands: Reidel. Reprinted in 1980 as a monograph by Harvard University Press.
Kripke. S. 1977. "Speaker's reference and semantic reference." In *Midwest Studies in Philosophy*, vol. 2, edited by P. French, T. Uehling, and H. Wettstein. Minneapolis: University of Minnesota Press.
Kripke. S. 1979. "A puzzle about belief." In *Meaning and Use*, edited by A. Margalit, pp. 239–283. Dordrecht, Netherlands: Reidel. Reprinted in Salmon and Soames, 1988, pp. 102–148.
Larson, R., and P. Ludlow. 1993. "Interpreted logical forms." *Synthese*, 95(3): 305–355.
Leeds, S. 1979. "Church's translation argument." *Canadian Journal of Philosophy*, 9(1): 43–51.
Lepore, E., and B. Loewer. 1989. "You can say *that* again." In *Midwest Studies in Philosophy*, vol. 14, edited by P. French, T. Uehling, and H. Wettstein, pp. 338–356. Minneapolis: University of Minnesota Press.
Lewis, D. 1972. "General semantics." In *Semantics for Natural Language*, edited by D. Davidson and G. Harman. Dordrecht, Netherlands: Reidel. Reprinted in Lewis, 1983, pp. 545–665.
Lewis, D. 1979. "Attitudes de dicto and de se." *Philosophical Review*, 88(4): 513–543. Reprinted in Lewis, 1983.
Lewis, D. 1983. *Collected Papers*, vol. 1. Oxford: Oxford University Press.
Linsky, L., ed. 1971. *Reference and Modality*. Oxford: Oxford University Press.
MacFarlane, J. 2014. *Assessment Sensitivity*. Oxford: Oxford University Press.
Martin, M. G. F. 1992. "Perception, concepts, and memory." *Philosophical Review*, 101(4): 745–763.
Mates, B. 1950. "Synonymity." *University of California Publications in Philosophy*, 25: 201–226. Reprinted in *Semantics and the Philosophy of Language*, edited by L. Linsky. Champaign, IL: University of Illinois Press.
McDowell, J. 1984. "*De re* senses." *Philosophical Quarterly*, 34(136): 283–294.
McGinn, C. 1983. *The Subjective View: Secondary Qualities and Indexical Thoughts*. Cambridge: Cambridge University Press.
McLaughlin, B. 1993. "The connectionism/classicism battle to win souls." *Philosophical Studies*, 71(2): 163–190.
Moltmann, F. 2014. "Propositions, attitudinal objects, and the distinction between actions and products." *Canadian Journal of Philosophy*, 43(5–6): 679–701.
Montague, R. 1974. *Formal Philosophy*, edited by R. Thomason. New Haven, CT: Yale University Press.

Neale, S. 1988. "Events and 'logical form.'" *Linguistics and Philosophy*, 11(3): 303–322.
Peacocke, C. 1981. "Demonstrative thoughts and psychological explanation." *Synthese*, 49(2): 187–217.
Pearson, H. 2012. "The Sense of Self: Topics in the Semantics of De Se Expressions." PhD diss., Massachusetts Institute of Technology.
Perry, J. 1979. "The problem of the essential indexical." *Noûs*, 13(1): 3–21. Reprinted in Salmon and Soames, 1988, pp. 83–101.
Perry, J. 1986. "Thought without representation." *Proceedings of the Aristotelian Society*, suppl. vol. 6: 263–283.
Perry, J., and M. Crimmins. 1989. "The prince and the phone booth: reporting puzzling beliefs." *Journal of Philosophy*, 86(12): 685–711.
Powers, L. 1978. "Knowledge by deduction." *Philosophical Review*, 87(3): 337–371.
Putnam, H. 1954. "Synonymity and the analysis of belief sentences." *Analysis*, 14(5): 114–122.
Putnam, H. 1975a. "The meaning of 'meaning.'" In Putnam, 1975b, pp. 215–271.
Putnam, H. 1975b. *Mind, Language, and Reality. Collected Papers*, vol. 2. Cambridge: Cambridge University Press.
Recanati, F. 2013. *Mental Files*. Oxford: Oxford University Press.
Richard, M. 1983. "Direct reference and ascriptions of belief." *Journal of Philosophical Logic*, 12(4): 425–452. Reprinted in Salmon and Soames, 1988, pp. 83–101, and Richard, 2013b, pp. 26–47.
Richard, M. 1987. "Attitude ascriptions, semantic theory, and pragmatic evidence." *Proceedings of the Aristotelian Society*, 87: 243–262. Reprinted in Richard, 2013b, pp. 65–79.
Richard, M. 1988. "Taking the Fregean seriously." In *Philosophical Analysis: A Defense by Example*, edited by D. Austin, pp. 219–240. Dordrecht, Netherlands: Reidel.
Richard, M. 1989. "How I say what you think." In *Midwest Studies in Philosophy,* vol. 14, edited by P. French, T. Uehling, and H. Wettstein, pp. 317–337. Notre Dame, IN: University of Notre Dame Press. Reprinted in Richard, 2013b.
Richard, M. 1990. *Propositional Attitudes*. Cambridge: Cambridge University Press.
Richard, M. 1993a. "Sense, necessity, and belief." *Philosophical Studies*, 69(2–3): 243–263. Reprinted in Richard, 2013b.
Richard, M. 1993b. "Attitudes and context." *Linguistics and Philosophy*, 16(2): 123–148. Reprinted in Richard, 2013b.
Richard, M. 1993c. "Attitudes, indexicality, and propositions." Series of lectures given at the National University of Mexico, Mexico City.
Richard, M. 1995a. "Propositional quantification." In *Logic and Reality*, edited by J. Copeland. Oxford: Oxford University Press. Reprinted in Richard, 2013b, pp. 137–155.
Richard, M. 1995b. "Defective contexts, accommodation, and normalization." *Canadian Journal of Philosophy*, 25(4): 551–570. Reprinted in Richard, 2013b.
Richard, M. 1997. "What does commonsense psychology tell us about meaning?" *Noûs*, 31(1): 87–114.
Richard, M. 2011. "Relativistic content and disagreement." *Philosophical Studies*, 156(3): 421–431.
Richard, M. 2013a. "What are propositions?" *Canadian Journal of Philosophy*, 43(5–6): 702–719. Reprinted in Richard, 2013b.
Richard, M. 2013b. *Meaning in Context,* vol. 1: *Context and the Attitudes*. Oxford: Oxford University Press.
Richard, M. 2013c. "Marcus on belief and belief in the impossible." *Theoria*, 28(3): 407–420.
Richard, M. 2015: *Meaning in Context,* vol. 2: *Truth and Truth Bearers*. Oxford: Oxford University Press.
Rosenthal, D., ed. 1991. *The Nature of Mind*. Oxford: Oxford University Press.
Russell, B. 1911. "Knowledge by acquaintance and knowledge by description." *Proceedings of the Aristotelian Society*, 11: 108–128. Reprinted in Salmon and Soames, 1988, pp. 16–32.
Russell, B. 1912. *The Problems of Philosophy*. Oxford: Oxford University Press.

Salmon, N. 1986. *Frege's Puzzle*. Cambridge, MA: MIT Press.
Salmon, N., and S. Soames, eds. 1988. *Propositions and Attitudes*. Oxford: Oxford University Press.
Saul, J. 1993. "Still an attitude problem." *Linguistics and Philosophy*, 16(4): 423–435.
Schiffer, S. 1979. "Naming and knowing." In *Midwest Perspectives in the Philosophy of Language*, vol. 2, edited by P. French, T. Uehling, and H. Wettstein, pp. 28–44. Minneapolis: University of Minnesota Press.
Schiffer, S. 1987. *Remnants of Meaning*. Cambridge, MA: MIT Press.
Schiffer, S. 1992. "Belief ascription." *Journal of Philosophy*, 89(10): 499–521.
Searle, J. 1990. "Consciousness, explanatory inversion, and cognitive science." *Behavioral and Brain Sciences*, 13(4): 585–598.
Segal, G. 1989. "A preference for sense and reference." *Journal of Philosophy*, 86(2): 73–89.
Sleigh, R. 1967. "On quantifying into epistemic contexts." *Noûs*, 1(1): 23–31.
Soames, S. 1984. "Lost innocence." *Linguistics and Philosophy*, 8(1): 59–71.
Soames, S. 1987. "Direct reference, propositional attitudes, and semantic content." *Philosophical Topics*, 15(1): 47–88. Reprinted in Salmon and Soames, 1988, pp. 197–239.
Soames, S. 1989. "Semantics and semantic competence." In *Philosophical Perspectives*, vol. 3, edited by J. Tomberlin, pp. 575–596. Atascadero, CA: Ridgeview Publishing Company.
Soames, S. 1992. "Truth, meaning, and understanding." *Philosophical Studies*, 65(1): 17–35.
Soames, S. 1995. "Beyond singular propositions?" *Canadian Journal of Philosophy*, 5(4): 515–549.
Soames, S. 2002. *Beyond Rigidity*. Oxford: Oxford University Press.
Soames, S. 2010a. "Coordination problems." *Philosophy and Phenomenological Research*, 81: 464–474.
Soames, S. 2010b. *What is Meaning?* Princeton, NJ: Princeton University Press.
Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, R. 1987. "Semantics for belief." *Philosophical Topics*, 15(1): 177–190.
Stalnaker, R. 1999. *Context and Content*. Oxford: Oxford University Press.
Stalnaker, R. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.
Stanley, J. 2011. *Know How*. Oxford: Oxford University Press.
Stich, S. 1978. "Beliefs and subdoxastic states." *Philosophy of Science*, 45(4): 499–518.
Stich, S. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Yagisawa, T. 1984. "The pseudo-Mates argument." *Philosophical Review*, 93(3): 407–419.

Further Reading

Anderson, C. A. 1983. "The paradox of the knower." *Journal of Philosophy*, 80(6): 338–355.
Anderson, C. A. 1987. "Review of Bealer's *Quality and Concept*." *Journal of Philosophical Logic*, 16: 115–164.
Anderson, C. A. 1989. "Russellian intensional logic." In Almog, Perry, and Wettstein, 1989, pp. 67–107.
Asher, N. 1986. "Belief in discourse representation theory." *Journal of Philosophical Logic*, 15(2): 127–189.
Bach, K. 1994. "Conversational implicature." *Mind & Language*, 9(2): 124–162.
Barwise, J. 1989. "Situations, facts, and true propositions." *The Situation in Logic*. Stanford, CA: Center for the Study of Language and Information.
Barwise, J., and J. Etchemendy. 1987. *The Liar*. Oxford: Oxford University Press.
Bealer, G. 1982. *Quality and Concept*. Oxford: Oxford University Press.
Burge, T. 1977. "Belief *de re*." *Journal of Philosophy*, 74(6): 338–362.
Burge, T. 1979. "Individualism and the mental." In *Midwest Studies in Philosophy*, vol. 4, edited by P. French, T. Uehling, and H. Wettstein, pp. 73–121. Minneapolis: University of Minnesota Press.
Burge, T. 1982. "Other bodies." In Woodfield, 1982, pp. 97–120.
Crimmins, M. 1992. "Context in the attitudes." *Linguistics and Philosophy*, 15(2): 185–198.
Cummins, R. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Dennett, D. 1981. *Brainstorms*. Cambridge, MA: MIT Press.
Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
Field, H. 1977. "Logic, meaning, and conceptual role." *Journal of Philosophy*, 74(7): 379–409.
Field, H. 1978. "Mental representation." *Erkenntnis*, 13: 9–61.
Fodor, J. 1981. *Representations*. Cambridge, MA: MIT Press.
Forbes, G. 1990. "The indispensability of Sinn." *Philosophical Review*, 99(4): 535–563.
Forbes, G. 1993. "Solving the iteration problem." *Linguistics and Philosophy*, 16(3): 311–330.
Frege, G. 1952. *Translations from the Philosophical Writings*, translated by M. Black and P. Geach. Oxford: Oxford University Press.
Frege, G. 1984. "Thoughts." In *Collected Papers on Mathematics, Logic, and Philosophy*, edited by B. McGuinness. Oxford: Blackwell. Reprinted in Salmon and Soames, 1988, pp. 33–55.
Higginbotham, J. 1991. "Belief and logical form." *Mind & Language*, 6(4): 344–369.
Higginbotham, J. 1993. *Language and Cognition*. Oxford: Blackwell.
Hintikka, J. 1962. *Knowledge and Belief*. Ithaca, NY: Cornell University Press.
Kaplan. D. 1986. "Opacity." In *The Philosophy of W. V. Quine*, edited by L. Hahn and P. Schlipp. La Salle, IL: Open Court.
Lewis, D. 1970. "How to define theoretical terms." *Journal of Philosophy*, 67(13): 427–446. Reprinted in Lewis, 1983.
Lewis, D. 1974. "Radical interpretation." *Synthese*, 23: 331–344. Reprinted in Lewis, 1983.
Linsky, L. 1985. *Oblique Contexts*. Chicago: University of Chicago Press.
Loar, B. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.
Loar, B. 1987. "Subjective intentionality." *Philosophical Topics*, 15(1): 89–124.
MacKay, T. 1979. "On proper names in belief ascriptions." *Philosophical Studies*, 39(3): 287–303.
MacKay, T. 1991. "Representing *de re* beliefs." *Linguistics and Philosophy*, 14(6): 711–739.
Marcus, R. B. 1981. "A proposed solution to a puzzle about belief." In *Midwest Studies in Philosophy*, vol. 6, edited by P. French, T. Uehling, and H. Wettstein, pp. 338–356. Minneapolis: University of Minnesota Press.
Marcus, R. B. 1983. "Rationality and believing the impossible." *Journal of Philosophy*, 80(6): 321–337.
McDowell, J. 1977. "On the sense and reference of a proper name." *Mind*, 86(342): 159–185. Reprinted in *Reference. Truth, and Reality*, edited by M. Platts. London: Routledge and Kegan Paul, 1980, pp. 141–166.
McGinn, C. 1982. "The structure of content." In Woodfield, 1982, pp. 207–258.
Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Moravcsik, J. 1990. *Thought and Language*. London: Routledge.
Morton, A. 1980. *Frames of Mind*. Oxford: Oxford University Press.
Parsons, T. 1981. "Frege's hierarchies of indirect senses and the paradox of analysis." In *Midwest Studies in Philosophy*, vol. 6, edited by P. French, T. Uehling, and H. Wettstein, pp. 338–356. Minneapolis: University of Minnesota Press.
Perry, J. 1977. "Frege on demonstratives." *Philosophical Review*, 86(4): 474–497.
Perry, J. 1980. "Belief and acceptance." In *Midwest Studies in Philosophy*, vol. 5, edited by P. French, T. Uehling, and H. Wettstein, pp. 338–356. Minneapolis: University of Minnesota Press.
Prior, A. N. 1971. *Objects of Thought*. Oxford: Oxford University Press.
Putnam, H. 1963. "Brains and behavior." In Putnam, 1975b, pp. 325–341.
Putnam, H. 1966. "The mental life of some machines." In Putnam, 1975b, pp. 408–428.
Putnam, H. 1967. "The nature of mental states." In Putnam, 1975b, pp. 429–440.
Putnam, H. 1970. "Is semantics possible?" In Putnam, 1975b, pp. 139–152.
Putnam, H. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.
Quine, W. V. O. 1956. "Quantifiers and propositional attitudes." *Journal of Philosophy*, 53: 177–187. Reprinted in Linsky, 1971.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Recanati, F. 1993. *Direct Reference*. Oxford: Blackwell.

Russell, B. 1903. *Principles of Mathematics*. London: Allen and Unwin.
Russell, B. 1956 (1918). "The philosophy of logical atomism." In *Logic and Knowledge*, edited by R. Marsh, pp. 175–281. London: George Allen and Unwin.
Salmon, N. 1986. "Reflexivity." *Notre Dame Journal of Formal Logic*, 27(3): 401–429. Reprinted in Salmon and Soames, 1988, pp. 240–274.
Salmon, N. 1989. "Illogical belief." In *Philosophical Perspectives*, vol. 3, edited by J. Tomberlin. Atascadero, CA: Ridgeview Publishing Company.
Salmon, N. 1992. "Reflections on reflexivity." *Linguistics and Philosophy*, 15(1): 53–63.
Schiffer, S. 1990a. "The mode-of-presentation problem." In *Propositional Attitudes*, edited by C. Anderson and J. Owens, pp. 249–268. Stanford, CA: CSLI Publications.
Schiffer, S. 1990b. "The relational theory of belief." *Pacific Philosophical Quarterly*, 71: 240–245.
Segal, G. 1990. "The return of the individual." *Mind*, 98: 39–57.
Soames, S. 1984. "What is a theory of truth?" *Journal of Philosophy*, 81(8): 411–429.
Soames, S. 1987. "Substitutivity." In *On Being and Saying: Essays for Richard Cartwright*, edited by J. J. Thompson. Cambridge, MA: MIT Press.
Soames, S. 1990. "Pronouns and propositional attitudes." *Proceedings of the Aristotelian Society*, 90: 191–212.
Stalnaker, R. 1976. "Propositions." In *Issues in the Philosophy of Language*, edited by A. MacKay and D. Merrill, pp. 79–91. New Haven, CT: Yale University Press.
Stalnaker, R. 1978. "Assertion." *Syntax and Semantics*, 9: 315–332.
Stalnaker, R. 1981. "Indexical belief." *Synthese*, 49(1): 129–151.
Stich, S. 1979. "Do animals have beliefs?" *Australasian Journal of Philosophy*, 57(1): 79–91.
Walton, K. 1990. *Mimesis as Make Believe*. Cambridge, MA: Harvard University Press.
Wettstein, H. 1991. *Has Semantics Rested Upon a Mistake?* Stanford, CA: Stanford University Press.
Woodfield, A., ed. 1982. *Thought and Object: Essays on Intentionality*. Oxford: Oxford University Press.

15

Holism

CHRISTOPHER PEACOCKE

The question must arise whether a doctrine which is attributed to all of Quine, Putnam, Davidson, Rorty, Gadamer, and Heidegger is possibly a doctrine which comes in more than one version. Even the most ardent taxonomist is likely to draw back from classifying the various actual and possible positions which emerge from the very tangled history of recent discussions of holism. I will be approaching the matter by addressing a series of questions, starting with those which are most likely to arise in the mind of those philosophers who regard holism with a mixture of fascination and suspicion.[1]

1 What Is Meaning Holism?

Here is a highly general formulation of global holism about meaning, a formulation acceptable to holists of many different stripes:

> (GH) The meaning of an expression depends constitutively on its relations to all other expressions in the language, where these relations may need to take account of such facts about the use of these other expressions as their relations to the non-linguistic world, to action, and to perception.

This is a constitutive thesis about what it is for an expression to have a certain meaning. It is neither an epistemological thesis, nor a psychological thesis; though of course if it is correct, it will have consequences for both psychology and epistemology. It goes far beyond the less controversial claim that in assessing the evidence that a given expression has a certain meaning, we must take account of the properties of any sentence in which the expression occurs, regardless of what else is in that sentence.

(GH) is non-committal in at least two respects. First, different theorists who are both committed to accepting (GH) may emphasize different relations to the non-linguistic world as partially constitutive of meaning. Some theorists accepting (GH) may give a special status to observable states of affairs; others may not. Empiricism is not written

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

into, nor entailed by, (GH). Second, one who holds (GH) is not committed to saying that one can make explicit, in a non-circular way, the relations to all other expressions in which a given expression must stand if it is to have a given meaning. Interpretationism is the doctrine that in saying what it is for a particular expression to have a given meaning, a fundamental place must be given to the fact that under optimal interpretation of the language in which the expression occurs, the expression is assigned that meaning. According to the interpretationist there may be more to be said about optimal interpretation in general, and more to be said about various particular meanings. But neither of these, according to the interpretationist, will amount to a full account of what it is to grasp a given meaning unless they actually mention optimal interpretation itself. This is not the place for a discussion of the important question of the correctness of interpretationism, and one of its possible motivations, a certain subjectivism about propositional attitudes and contents (as in McDowell, 1986, and possibly Davidson himself). All that matters for present purposes is that it is *prima facie* consistent for a global holist about meaning to be an interpretationist. (For discussion of Davidson's position, see Chapter 13, RADICAL INTERPRETATION.)

Global holists who, unlike interpretationists, believe that grasp of a particular meaning can be made explicit without presupposing the understander's grasp of that meaning, can make use of the notion of an *understanding-condition* in specifying meanings. An understanding-condition is an explicit statement of the condition a person must meet to understand a given expression, a statement which does not at any point take for granted understanding of the expression in question, nor possession of the concept it expresses. We can formulate the global holism of these theorists in:

(GHE) For a thinker to meet the understanding-condition for any expression E, there must exist certain other expressions $E_1, \ldots, E_n$, such that

(a) the understander meets a certain specifiable, non-circular condition $R(E, E_1, \ldots, E_n)$; this condition may concern the use of the expressions $E_1, \ldots, E_n$, and it may concern their relations to the non-linguistic world; and

(b) the expressions E, E_1, …, E_n exhaust the expressions in the understander's language.

Global holism of the sort captured in (GHE) itself comes in several kinds. One kind is that variety which recognizes certain methods of establishing sentences containing a given expression, or certain methods of deriving consequences from them, as canonical. It writes these methods into the relevant understanding-conditions. One example of a global holism with canonical methods, (GHEC), restricted to the language of mathematics, would be the "pure mathematical holism" mentioned by Dummett, which identifies truth with provability by any of the canonical methods acknowledged in classical mathematics (Dummett, 1991, p. 226). Since some sentences can be proved outright by classical methods, we have in this example one of the extreme forms of holism attacked by Dummett, according to which understanding another person's expression sometimes involves knowing which sentences containing it he holds true (Dummett, 1975, appendix). This framework of classification also formally leaves space for a further kind of case, that of a kind of holism which does not accept the designation of any methods as canonical (GHENC). Such, for instance, would be the position of a global holist who thinks that we can usefully speak only of similarity of

meaning, not of identity of meaning, and who holds that two sentences are more similar in meaning the greater the overlap in accepted methods of establishing them.

Those already well acquainted with this territory will recognize that (GHE) is the natural formulation of that kind of global holism when the apparatus of possession conditions, as a means of individuating concepts, is adapted to the case of linguistic meaning (Peacocke, 1992). I have chosen this formulation for its bearing on the issue of the circularity of global holism. Those who reject all forms of global holism, but still accept the possibility of explicit formulations of understanding-conditions, commonly propose that there is a certain non-trivial, partial ordering of all the expressions of a language. This ordering has the property that to elucidate the meaning of an expression at any given place in the ordering, it is not necessary to mention its relations to expressions later in the ordering. Since the global holist will say that there is no such ordering, it is natural to wonder whether a global holist can avoid circularity in his account of meaning. Dummett, for one, characterizes holism as "the doctrine that any meaning-theory is inevitably circular" (1991, p. 241). In fact the availability of the above form (GHE) should indicate that there is no structural obstacle of principle to the global holist's giving non-circular accounts of the understanding of particular expressions. For instance, one form of global holism might state that to understand a certain expression, one must appreciate that sentences containing it can be established by a certain finite family of methods, where a statement of the methods involves all the other concepts expressed in the language. This may not be plausible – it will be discussed below – but provided that the methods can be specified without presuming on any prior understanding of the expression in question, it is not circular. A global holist who possesses such non-circular specifications may rightly insist that he admits and employs the notion of the content of a given individual sentence. For him, that content will be fixed by the meaning specifications for the components of the sentence, together with the way in which they compose the sentence.

Quine is not a holist of any of the sorts so far distinguished, for he has always acknowledged a level of observational vocabulary for which he would say that all of the above theses are false. Quine's holism is captured by the preceding formulations only if we understand the talk of expressions, and variables ranging over them, as restricted to the non-observational part of the language. We are clearly entering here the realm in which holism comes in degrees. A position may be classified as more holistic the fewer restrictions we have to place on the talk of expressions in order for the position to be classified as holistic according to the characterizations above. This matter of degree arises also for conceptual-role theories of meaning, as advocated in Block (1986), Harman (1982), and Sellars (1974), which state that for an expression to have a particular meaning is for it to have a certain role in its user's psychology. A conceptual-role theory of meaning will similarly be more holistic the fewer restrictions it places on those features of an expression's total role which it regards as individuative of the expression's meaning. The limiting case at the end of the spectrum of increasingly holistic alternatives is that in which no restrictions are placed on those features of the conceptual role of an expression which contribute to individuating its meaning. Acceptance of this limiting case is naturally accompanied by skepticism that the strict relation of intersubjective synonymy of expressions – at least for that aspect of meaning captured by holistic conceptual role – is ever, in fact, satisfied (cp. Field, 1977).

I turn now to consider grounds which have been offered in support of meaning holism.

2 Does the Duhem–Quine Thesis Provide a Ground for Meaning Holism?

This question must be split into two parts:

(a) Is the Duhem–Quine Thesis true? (If it is not, it will not be a ground for anything.)
(b) If the Duhem–Quine Thesis is true, does it support meaning holism?

Actually we should distinguish the Duhem thesis from the Quine thesis. Duhem wrote:

> the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (Duhem, 1962, p. 187)

Two points emerge from Duhem's discussion. First, Duhem's thesis is specific to hypotheses of physics. He explicitly contrasted the physicist's situation with that of a physiologist who wishes to confirm that a nerve is a motor nerve, rather than a sensory nerve (Duhem, 1962, p. 182). Second, in Duhem's account it is the experiments which are said to confirm, or to be in conflict with, a group of hypotheses. Quine's thesis, which he says "was well argued by Duhem" (Quine, 1961, p. 41, n. 17), is by contrast not confined to physics; and for Quine what confirms or conflicts with groups of hypotheses are not the results of experiments but rather (in the 1951 version) sense experiences.[2] Quine wrote:

> our statements about the external world face the tribunal of sense experience not individually but only as a corporate body. (Quine, 1961, p. 41)

And, further on in the same paper:

> Even a statement very close to the periphery [of our field of beliefs] can be held true in the face of recalcitrant experience by pleading hallucination. (Quine, 1961, p. 43)

Duhem could not have offered that ground on behalf of *his* thesis. The possibility of pleading hallucination arises no less for the physiologist than for the physicist. A thesis defensible on such grounds could not discriminate between those two disciplines.

In Quine's later formulations, the talk of sense experiences gives way to talk of stimuli. With this comes his notion of the stimulus-meaning of a sentence for a speaker, which is the ordered pair of its affirmative and negative stimulus-meanings. Its affirmative stimulus-meaning is "the class of all stimulations (hence evolving ocular irradiation patterns between properly timed blindfoldings) that would prompt his assent" (Quine, 1960, p. 32). Negative stimulus-meaning is defined similarly, with "dissent" in place of "assent." In this later framework, the Quine thesis becomes the claim that sentences about the external world cannot be assigned stimulus-meanings one-by-one, but only collectively, in sets.

This version of the Quine thesis is plausible, but it supports holism about meaning only if meaning is to be elucidated in terms of stimulus-meaning. Identity of stimulus-meaning is far from necessary for identity of meaning. Creatures with very different sensory systems

could mean the same thing, on an occasion, by an utterance of the sentence "This edge is straight," even though their different patterns of sensory receptors preclude identity of stimulus-meaning. Nor do the sensory systems have to be radically different for the point to hold. For someone who knowingly has a serious case of astigmatism, the stimulus-meaning of "That line is straight" will differ from its stimulus-meaning for his better-sighted friend. Yet the sentences have the same meaning for both. The lesson of such simple examples is twofold. Meaning must be keyed more strongly to the environment; and we cannot hope to capture the nature of a person's grasp of meaning by looking solely at incoming information, to the neglect of the person's later use of that information, including ultimately its effects on his actions. Invited on many occasions to endorse a firmer separation of meaning from stimulus-meaning, Quine has persistently held fast to his later formulations: "I did intend the stimulus meaning to capture the notion of meaning – for the linguistic community in the case of an observation sentence, and for the individual speaker in the case of many other occasion sentences" (Quine, 1986, pp. 427–428).

Does Duhem's thesis imply a form of meaning holism? We can formulate the crucial question more generally, and also more explicitly. Let us say that a given branch of discourse *has the Duhemian property* just in case it is only whole groups of statements in that branch of discourse which are confirmed by experiments or observable states of affairs, rather than individual statements. So a branch of discourse is Duhemian if and only if it is as Duhem thought physics to be. We can now formulate our question thus: Are the distinctive terms of a Duhemian branch of discourse such that their understanding-conditions involve all the other terms of the whole language?

There are two reasons we should not give an unrestricted affirmative answer. The first is that there are some examples of areas of discourse which have the Duhemian property, but where the most plausible explanation of the phenomenon does not involve any global holism about the meaning of its distinctive terms. One example is discourse about persons' intentional states, and the actions they explain. Any particular action may potentially be explained by indefinitely many combinations of beliefs and desires. Even if one is given in advance both that a particular event is an intentional action, and the description under which it is intentional, nothing follows about which mental states of the person whose action it is were operative in producing the action. In general, actions are produced only by combinations of beliefs and desires. Correlatively, their occurrence can confirm or disconfirm only whole sets of hypotheses about the agent's mental states. There is disagreement about what grasp of the scheme of explanation of intentional states involves – whether it involves approximate laws, or whether it involves some irreducible notion of making something intelligible, to mention two of the options. But it cannot be plausibly suggested that the concepts of all the other sciences have to be brought in to explain what is involved in mastery of the scheme of intentional states. Any evidence may, indeed, be relevant to such questions as that of whether certain normal conditions, which may be required for perception, or reasoning, or intentional action are met; but this is holism of the evidence, not meaning holism.

What holds for intentional psychology may also hold for other areas of discourse. The language a person employs may be divisible into many different parts, and the discourse of each part may have the Duhemian property, without any form of global meaning holism being true of the language as a whole.

The second reason against saying that any Duhemian branch of discourse supports a form of global holism is more fundamental, and indeed, if correct, suggests that the first

reason may already be implicitly conceding too much. Why should the meaning of a theoretical hypothesis, or of a set of them, be elucidated in terms of its or their consequences for observable states of affairs at all? A scientist may formulate this hypothesis: "There are particles of matter less than 0.000001 mm in diameter, which exert tiny forces on each other." It is initially quite implausible that in order to understand this hypothesis, the scientist must know observational tests for sets of hypotheses containing it. On the contrary, attaining knowledge of observational tests for this or any other hypothesis takes reasoning and creative thought. Such knowledge usually comes after understanding of the hypothesis, and so cannot be identified with the understanding. The capacity for understanding the hypothesis is present as soon as the thinker has a general notion of size, its measurement, of matter, and of forces. This understanding must indeed ultimately connect the measurement of size at various points with the observable, and a detailed account of the nature of that connection should be given. But it is one thing to state that such connections must exist for anyone who understands measurement; it is quite another to say that the meaning of a set of hypotheses is given in terms of their observational consequences. Grasp of systems of measurement for the various physical magnitudes mentioned in the special theory of relativity no doubt involves some indirect connection of values of these magnitudes with observables. But it took further thought to devise an observational test of the theory.

From the standpoint of this second reason, the first reason implicitly conceded too much in not contesting the claim, even for the non-global holism of the scheme of intentional explanation, that the meaning of its hypotheses are given by connections with the actions they explain. It can equally take creative thought to reason out what actions would be evidence, in the context of others a subject possesses, of a given propositional attitude. In this case, too, knowledge of what would be evidence is subsequent to understanding of the sentence which attributes the given attitude. Knowledge of what would be evidence cannot be identified with understanding that sentence.

3 Does Revisability Support Meaning Holism?

The rational revisability of statements has loomed large in discussions of holism. In "Two dogmas of empiricism," Quine argued that on the conception of empirical content he presents there, "no statement is immune to revision" (Quine, 1961, p. 43). In Putnam's discussion of "the considerations that lead *me* to embrace meaning holism," he highlights the rational, non-stipulative revisability of a vast range of statements (Putnam, 1986, pp. 406 ff.). I turn now to consider the relations between revisability and holism; and also to distinguish two rather different sources of revisability.

Meaning holism does not by itself imply unlimited revisability. Those forms of meaning holism which admit certain canonical methods – the forms which accept (GHEC) of §1 above – actually preclude unlimited revisability, since according to them suitable uses of the canonical methods will not be revisable. The same holds for more limited forms of holism which admit canonical methods. Canonical methods are in effect acknowledged in Quine's later thought, in which those principles to which assent is ensured by his 'verdict tables' are taken as having an innocent kind of analyticity (Quine, 1974, pp. 77–80). Only those forms of meaning holism which do not distinguish any methods at all as canonical entail unlimited revisability. It is a real question whether these forms can make any sense of the immediate acceptance of certain statements as being required by reason, and of certain transitions

as being required by reason. Some of these forms are certainly excluded by the fact that we could never rationally revise statements of the form "If *p*, then *p*."

Does Duhem's thesis imply some kind of revisability which in its turn supports meaning holism? It cannot, if the considerations of the previous section were correct. I argued that Duhem's thesis does not imply meaning holism. If that is right, Duhem's thesis can hardly imply something which in turn implies meaning holism. But it is well worth considering what kinds of revisability are present in Duhemian areas of discourse (in the sense of the preceding section), in order to distinguish them from examples of revisability which have a different source.

A paradigm case of revisability of the sort which impressed Duhem is the attribution of a particular numerical value to a theoretical magnitude, ascribed to a particular object (or region) at a particular time. When this sort of attribution is made on the basis of experimental data, it seems indisputable that the attribution would have to be revised if, as could well be the case, we came rationally to change our mind about the principles linking the theoretical magnitude with the observable properties of the experimental setup. But besides being no challenge to central laws of logic, revisability of such particular ascriptions of theoretical magnitudes to particular objects is also consistent with the unrevisability of statements of certain very general characteristics of theoretical magnitudes or properties. A magnitude that has theoretical links with neither mass nor repulsion and attraction could hardly be the magnitude of force. Something which has nothing to do with heritable characteristics could hardly be a gene. Duhem's thesis does not exclude such simple examples of unrevisability.

Revisability with a rather different source can be illustrated if we take, first, perceptual demonstratives. Consider the perceptual demonstrative "that plant," which is in a suitable context a way of thinking of a particular plant made available by the plant's being presented to the thinker in perception in a particular way. The thinker may radically revise his view of various statements of the form "that plant is thus-and-so" without "that plant" losing its reference nor, more importantly for our concerns, its sense. These radical revisions may involve change of belief about the plant's origins, its sources of energy, its mode of reproduction, its lack (or otherwise) of magical properties, and much else. Revision on any of these matters does not affect the sense of the demonstrative expression, because no such revision undermines the foundation of the availability of the perceptual-demonstrative way of thinking of the object, *viz*. its being, in virtue of its causal relations to his perceptual state, the one which is presented to him in a certain way in perception. More generally, we can describe a kind of sense (or mode of presentation) as *causally linked* if a correct statement of what is required for something to be the reference of a sense of that kind mentions the causal relations of that thing to the thinker. Perceptual demonstratives constitute just one of many causally linked kinds of sense. Another type is that of recognitionally based senses; a further kind is that whose instances have their references fixed in part by their being the dominant sources of certain dossiers of information.[3] In these, as in other causally linked cases, recognition of radical error, and hence radical revision of beliefs, is consistent with constancy of sense. Many of Putnam's most striking examples of revisability are ones which turn on a causally linked way of thinking of some property or magnitude (Putnam, 1986).

Rational revisability made possible by causally linked senses does not seem to me to give any grounds for meaning holism in the sense distinguished in §1 above. Examples of such rational revisability are consistent with the predicates whose ascription is revised having understanding-conditions which do not make reference to the whole language. Nor is it

plausible that the understanding-conditions of the expressions with causally linked senses presuppose understanding of all the rest of the language. The crucial element in the understanding-conditions for those expressions is that the thinker's use of them be suitably answerable to information coming via the causal channel or channels which make available to the thinker the example of the causally linked kind in question. A thinker with only a fairly primitive vocabulary and conceptual repertoire can be making use of the same instance of a causally linked kind as a powerful theoretician.

Quine's arguments for holism and his insistence on extensive revisability have been linked in his writings with a rejection of analyticity, understood as truth "purely in virtue of meaning." If we reject the holism, and assert that revisability is more limited than Quine allows, are we thereby committed to the existence of sentences true purely in virtue of meaning? We are not. It is entirely open to one who rejects holisms of a Quinean sort to agree nonetheless that no sense can be made of truth purely in virtue of meaning. Any sentence, if true, can be said to be true in virtue of its disquoted truth-condition, as Quine himself has long insisted. From the standpoint of more recent approaches to meaning, the possibility of any sentence's being "true purely by virtue of meaning" remains highly questionable. Suppose we have a truth-conditional approach to meaning. Then for each of the expressions in a sentence, its meaning is given by its axiom in that theory of truth which can serve as a meaning-theory for the language. What could being "true purely in virtue of meaning" amount to in this framework? The best candidate would be that the truth of a sentence so classified is derivable solely from the axioms of the truth theory dealing with the individual components of the sentence. But in that case, it is certain that no sentences are true purely in virtue of meaning, because the meaning-theoretic axioms have to be supplemented with logic if theorems are to be derived from them. Under this elaboration of "truth in virtue of meaning," a logical truth can be shown to be true in virtue of the meaning of its component, *plus* logic. This does not look like an alternative to the principle that every sentence is true in virtue of its disquoted truth-condition; rather, it looks like a special case of it. Indeed, the most obvious way to derive the outright truth of a logical truth in the truth theory is first to derive a T-sentence for it; then to prove outright, using logic alone, that the right-hand side of the T-sentence holds; and then to apply modus ponens right-to-left on the biconditional which constitutes the T-sentence. All this amounts to is showing that the sentence has a (canonical) truth-condition, and that its truth-condition holds.

What does, arguably, come with limited revisability and canonical methods is not analyticity, but rather a form of the *a priori*. It is plausible that semantic values are assigned to expressions in such a way that canonical methods involving those expressions are always truth-preserving. If this is correct, application of a canonical method will be truth-preserving, however the actual world may turn out to be. This sounds very close to a traditional form of the *a priori* (cp. Peacocke, 1993a). But it does not resuscitate truth-purely-in-virtue-of-meaning. A conditional sentence whose antecedent captures the input, and whose consequent expresses the output, of a canonical method can be said to be *a priori*; but it is still true in virtue of the holding of its disquoted truth-condition. Similarly, a canonical form of inference is truth-preserving because every instance is such that, if the truth-conditions of its premises are met, so are those of its conclusion. Recognition of canonical methods and a rejection of holism about meaning does not require one to side with Carnap on the possibility, or even the intelligibility, of truth purely by convention.[4] (For further relevant discussion, see Chapter 23, ANALYTICITY.)

4 Do Interpretational and Compositional Considerations Support Meaning Holism?

In a famous paper, "Truth and meaning," Davidson wrote:

> If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (and word) in the language. (Davidson, 1984, p. 22)

This is naturally read as a statement of global holism about meaning. But when considering the doctrine that the meaning of an expression is understood as an abstraction from the totality of sentences in which it occurs, we need to distinguish a constitutive from an epistemological version. The constitutive version states that *what it is* for an expression to have a certain meaning is to be explained by mentioning what can be abstracted from properties of whole sentences in which it occurs. It may help to draw a parallel with another constitutive thesis of abstraction. It is very tempting to hold that what it is for a given number to measure a particular object's mass (or other physical magnitude) is to be explained by mentioning the way in which that number simply codes a certain place that the object has in a system of physical relations which do not involve numbers. The numerical value of the physical magnitude does no more than abstract from a certain place in a system of relations. Representation theorems of the sort proved in the theory of measurement then allow us to say this: for an object to have a certain number as its mass is simply for it to be mapped to that number by the unique mass function which (a) takes a certain object as the unit mass, and (b) conforms to the two principles that the mass of x = the mass of y iff x has-the-same-mass as y, and that for non-overlapping objects, the mass of their sum is the sum of their masses. The constitutive thesis of abstraction for numerical values of physical magnitudes is very attractive. We seem completely unable to offer any constitutive account of what it is for an object to have, say, a mass of 5 grams which does not mention such abstraction. This is to agree with Field in denying what he calls 'heavy-duty Platonism' (Field, 1989, pp. 186–193).

The weaker, epistemological version of the thesis of abstraction for the case of meaning agrees that in coming to know the meaning of an expression we do not understand, evidence from the use of any sentence containing the expression may be relevant. But this point about evidence does not support global holism about meaning. Rather, evidence about any sentence containing an expression may be relevant to learning which one of several non-globally individuated meanings it possesses. Considerations in favor of the stronger, constitutive thesis must go beyond those which could equally be accommodated by the epistemological thesis.

There are two closely related problems for the holistic, constitutive thesis of abstraction in the case of meaning. The first problem is whether there exists any meaning-free level of properties and relations of sentences from which their meanings can be abstracted without dependence on meaning-involving notions. It is true that in Davidson's earlier accounts of radical interpretation, the fundamental level of evidence available to a radical interpreter was said to be that of holding a sentence true, an attitude which, it was emphasized, can be known to be present without knowing what the sentence means. In the earlier work, the constraint on a theory of truth as providing an interpretation of the language was said to be that of maximizing true belief, under the theory of truth in question (the "Principle of

Charity"). This constraint, too, is stated without attributing, or making hypotheses about, particular meanings. Davidson's later writings formulate the constraint on acceptable interpretations as that of maximizing intelligibility. As Grandy (1973) and McGinn (1977) emphasized, this diverges from the Principle of Charity. It is acceptable to attribute intelligible error; it is unacceptable to attribute inexplicably correct belief. Fulfillment of the constraint of maximizing intelligibility must involve the fulfillment of constraints for each particular content *p* which may be judged. If the belief that *p* is ascribed under an interpretation of the language-user's linguistic and non-linguistic behavior, then the language-user's behavior and attributed attitudes must be intelligible in the light of his believing the particular content that *p*. This is disanalogous to the relation between numerical values of physical properties and the underlying non-numerical physical relations from which they are abstracted. (It is as if there were particular constraints relating to an object's having a mass of 5 grams which any assignments of mass had to satisfy! See further Chapter 13, RADICAL INTERPRETATION.) In particular, this constraint on ascriptions of beliefs that *p*, and the general constraint of maximizing intelligibility, simply use intentional notions. They are not meaning-free accounts of how constitutive abstractionism might be true.

This difference between a constitutive abstractive thesis about meaning and other constitutive abstractive theses need not be fatal if the constraints relating to particular contents could be elucidated without taking for granted the notion of meaning or content. One could conceive, for example, of an alliance between the constitutive abstractionist about meaning and a conceptual-role theorist of meaning, one who insists that meaning is captured only by global conceptual role. This leads us, though, to the second problem.

Accounts have gradually been emerging of what is involved in mastering various particular concepts – in particular demonstratives (Evans, 1982), logical concepts, and observational concepts (Peacocke, 1992). These accounts have not adverted to the global conceptual role of the concept treated. They have, rather, concerned certain canonical circumstances for applying the concept, and certain canonical commitments involved in applying it. If these accounts are, even in principle, along the right lines, then it is not true that the meanings of words expressing these concepts are constitutively dependent upon properties of all the sentences in which they occur. For one who accepts such accounts, it is their existence which makes the case of meaning unlike that of the assignment of numbers as values of physical magnitudes (at least in the respect we have been discussing).

Interpretationists and others would, of course, doubt that these accounts can be completely correct if they purport to exhaust what individuates a meaning or concept. But this doubt need not push an interpretationist in the direction of global holism. It is open to the interpretationist to say that what makes it the case that a word expresses a particular observational concept is dependent upon the intelligibility of a person's use of the expression in certain perceptual circumstances, together with what he is intelligibly prepared to infer from the applicability of the expression in other circumstances. Similarly, the doubts of non-interpretationist objectors about the exhaustiveness of the offered accounts are more likely to concern their (somewhat) reductive features than their failure to embrace global holism.

It is important to note two points about the position which accepts the merely epistemological version of the abstractionist claim, while rejecting the stronger constitutive, holistic version. First, the merely epistemological version can still endorse the view that interpretation is answerable to global constraints of rational intelligibility. The core of the intelligibility requirements for each meaning attributed can be given by the thinker's satisfaction of the non-holistic possession condition for the concept expressed by an expression with that

meaning. The totality of such requirements for each meaning attributed gives a global set of requirements for an interpretation, a set answerable to the interpretee's use of any sentence in his language. The non-holistic interpretationist is then equally entitled to make the point that interpretation is answerable to global constraints of rational intelligibility. The global background constraint of maximizing intelligibility as applied to non-linguistic as well as linguistic actions is also applicable throughout the enterprise of interpretation. It, too, is independent of holism about meaning.

The second point to note is that the merely epistemological version of the abstraction claim can insist that an expression has linguistic meaning only if it is capable of combining with others to form complete sentences (when it is not already a complete sentence), and that this is so as a constitutive matter. Such an uncontroversial doctrine does not entail global holism about meaning. It also neither entails nor precludes less extensive holisms.

Reference and satisfaction are thought of very differently under the constitutive and the merely epistemological versions of the abstraction claim. Under the constitutive, holistic version, the correctness of attributions of referential relations to terms and of satisfaction conditions to predicates is exhausted by their role in contributing to the truth-conditions ascribed to individual sentences by a theory of truth. Consistently with his constitutive version of the abstractionist claim, this is precisely Davidson's position:

> these notions [satisfaction, reference – CP] we must treat as theoretical constructs whose function is exhausted in stating the truth conditions for sentences. … A theory of this kind … assigns no empirical content directly to relations between names or predicates and objects. These relations are given a content *indirectly* when the T-sentences are. … [Reference] plays no essential role in explaining the relation between language and reality. (Davidson, 1984, pp. 223, 225)

On Davidson's view, a particular truth theory "can be supported by relating T-sentences, and nothing else, to the evidence" (Davidson, 1984, p. 223).

There are at least three possible views about the role of reference in the explanation of facts about the truth-conditions of whole sentences. It seems to me that the correct view is intermediate between two extremes. At one extreme, we have the view which treats the role of reference as entirely analogous to the role of microproperties and microentities in the physical explanation of macrophenomena. If we have a realistic attitude to these physical theories, we will insist that the content of a statement about microphenomena is not exhausted by its role in the theory. We will also insist that it is neither constitutive nor necessary of microphenomena that they play that role in the explanation of macrophenomena. The macrophenomena might not even exist, consistently with the existence of the microphenomena and entities appealed to. The properties of carbon atoms explain the macroproperties of diamonds, but carbon atoms with their microproperties could still exist even were there not to be any diamonds. Davidson does in fact compare his own view of the status of the relation of reference, as expressed in the passages quoted in the preceding paragraph, with the status of the postulation of a fine structure in physical phenomena which explains macrophenomena (Davidson, 1984, p. 222). But the points just made about a realistic attitude to physical theories seem to me to show that a parallel with physical theory could be accepted by someone with a Davidsonian attitude to reference only if he took some form of instrumentalistic attitude to statements about microphenomena.

It does not seem correct simply to assimilate the role of axioms of reference in a semantic theory to axioms postulating microproperties and microentities in a physical theory. It should be agreed, even by the merely epistemological abstractionist, that, as an *a priori* and constitutive matter, the correctness of an axiom stating the reference of an atomic expression in a language is answerable to facts about complete sentences containing the expression. Truths about the relation of reference for atomic expressions cannot have the same metaphysical, constitutive independence of facts about the truth-conditions of complete sentences which microphenomena have from (at least certain) macrophenomena. It certainly seems that we cannot make sense of an atomic expression having a certain reference except in so far as its doing so contributes to the semantic properties of complete sentences in which it occurs. For this reason, there is at least one respect in which the view of reference which accompanies the merely epistemological version of the abstractionist doctrine need not be treating the concept of reference "as a concept to be given an independent analysis or interpretation in terms of non-linguistic concepts" – which is what Davidson was concerned to avoid (Davidson, 1984, p. 219).

At the other extreme we have the view that if the T-sentences of two semantic theories are the same, and are both well confirmed, then the two theories are equally good. This is an extreme form of instrumentalism about the semantic properties of sub-sentential expressions. This view has been vigorously contested – refuted, it seems to me – in the literature on tacit knowledge (see for instance Davies, 1987; Evans, 1981). The arguments are discussed elsewhere in this volume (see Chapter 12, TACIT KNOWLEDGE). What is important here is that there is a middle position between extreme instrumentalism and the view that treats semantic theory as analogous to physical theory. Claims about the reference of an atomic expression are constitutively and *a priori* answerable to facts about whole sentences containing it. But when a particular referential axiom for an expression *a* is correct, the explanation of a person's use and understanding of sentences containing *a* has a certain structure. Suppose a proposed axiom states that *a* denotes Paris. If speakers of the language happily assert all sorts of sentences of the form "*a* is thus-and-so" without checking on anything about Paris, and without taking themselves as answerable to anything about Paris, that is strong evidence against the proposed axiom. Generally, in the enterprise of maximizing intelligibility, a proposed axiom which states that an expression refers to a certain object is answerable to the role of the properties of that object in the explanation of speakers' sincere assertions containing the expression. More specifically, when the axiom "*a* denotes Paris" is correct for a language as understood by a given person, there is a common component in the explanation of all the various cases of his understanding of a sentence containing *a* as meaning something about Paris. The common component of each explanation is just his possession of the information stated in the axiom "*a* denotes Paris."

There is a great deal more to be said on many aspects of this middle position, but I hope enough has been said for us to be able to identify the two properties which distinguish it from the two extremes. On the middle position, it will be agreed that there is a sense in which it is *a priori* that: if *a* denotes Paris, and *f* is true of anything just in case it is elegant, then *fa* is true iff Paris is elegant. A full statement of the *a priori*, constitutive features of the relation of reference precludes assimilation of its status to something analogous to that of theoretically postulated relations in physics. But these *a priori* links in no way rule out the possibility that a person's possession of the information that *a* denotes Paris contributes causally to the explanation of his knowledge that *fa* means that Paris is elegant. There is such explanation when understanding is suitably structured; and so the middle position is also distinguished from the other, instrumentalist, extreme.

5 Global Holism, Justification, and Semantic Value

Dummett describes the holist thus:

> For the holist, we ought not to strive to command a clear view of the working of our language, because there is no clear view to be had. We have a haphazard assembly of conventions and rules, and there are no principles which govern our selection of them or render them any more appropriate than any others we might adopt. (Dummett, 1991, p. 241)

But we should, Dummett holds, subject any language to a critical scrutiny which

> aims at a systematic means of ascribing *content* to the expressions and sentences of the language, in terms of which accepted modes of operating with it (including the rules of inference observed) can be justified, or, better, are evidently justified. (Dummett, 1991, p. 241)

It was not written into our original formulations of holism – (GH), (GHE), and their variants – that such justification is unavailable. We also noted, towards the end of §1, that certain types of holist could endorse the notion of the content of an individual sentence. So we should consider separately types of holism which do, and types of holism which do not, make the claim that content-based justification is impossible. Let a *warranting* form of holism about meaning be a form which meets these two conditions. (1) The form of holism is committed to holding that there is a notion of justification on which certain assertions, made in appropriate circumstances, are warranted, in part by virtue of their meaning (and similarly for certain transitions between sentences). (2) This relation of justification is sufficiently powerful to rule out certain otherwise apparently satisfactory, fundamental specifications of alleged meanings (or understanding-conditions) as legitimate. We can say that a form of holism is *warrant-free* if it is committed to holding that there is no notion of justification meeting the conditions (1) and (2). It is a warrant-free holism which "sanctions the claim that we have a right to adopt whatever logical laws we choose" (Dummett, 1991, p. 227).

A warranting holism may be either conservative or revisionary of our actual judgmental and inferential practices, according as actual practices do or do not meet the standard of justification favored by the warranting holism. The scope for a revisionary warrant-free holism is much narrower. Some such scope no doubt exists. For example, a warrant-free holist may also believe in a form which classifies some methods of forming judgments as canonical; so the actual inferential practice of an individual may be criticized as not properly related to the canonical methods for the communal language. But this is a very limited kind of case. What warrant-free holists cannot coherently do is to criticize a practice as not meeting the requirements for justification, where justification is of the sort mentioned in conditions (1) and (2).

Our question must now be: Is either warranting holism or warrant-free holism about meaning tenable?

Warrant-free holism faces the problem of the existence of rules for certain expressions which lack meaning. These rules are in no way circular or infinitely regressive; but they fail to determine a meaning for the expressions they treat. Intuitively, they fail precisely because what is said in the proposed rules makes it impossible to see what the contribution of these expressions to the content of complete sentences containing them could possibly be. The

most well-known and spectacular case of this is the example of Prior's connective *tonk*, whose alleged sense is introduced by the two rules that from *A* one can infer *AtonkB*, and from *AtonkB* one can infer *B* (Prior, 1960). The rules for *tonk*, if accepted, clearly lead to the provability of all formulae. There are, though, other examples in which a proposed set of "rules" is not inconsistent (nor leads to radically non-conservative extensions of systems to which it is added), but fails to determine a meaning.

In earlier work, I mentioned the example of a spurious quantifier *Q* (Peacocke, 1993b). *Q* is said to have the same introduction rule as the existential quantifier: from $A(t)$, one can infer $QxA(x)$, subject to the usual restriction on the variables. *Q* is also said not to have any other introduction rules, and it is further said that the analogue of the existential elimination rule is invalid for it. There is a powerful intuition that no meaning is fixed by these rules. What could $QxA(x)$ possibly mean? It must mean something that can be inferred from any instance. Yet it cannot mean the existential quantification of $A(x)$, otherwise the elimination rule would be valid. It cannot mean something equivalent to the holding of an alternation consisting of that existential quantification with some further condition *p*. For then there should be a further introduction rule, that $QxA(x)$ can be inferred from *p* – yet there were said to be no further underived introduction rules. This argument that *Q* has no meaning is quite general, and applies under both constructivist and more realistic conceptions of content. The example is one of many. Similar points could be elaborated for the equally problematic connective "*Ü*," which was introduced and exposed by Dummett (1991, pp. 288–290). *Ü* is supposed to have the same introduction rule as ordinary alternation, but a more restricted elimination rule. The reader can develop an argument entirely parallel to that given for *Q* that there can be no contribution to the meaning of complete sentences containing it made by *Ü*.

It is highly plausible that what is wrong with all of these spurious connectives and operators is that the specifications placed upon them prevent them from having semantic values, appropriately related to their specifications, which contribute to the determination of the truth-value of complete sentences in which they occur. The rules for *tonk*, for example, place inconsistent requirements upon the truth-value of *AtonkB*, for the line of the truth table in which *A* is true and *B* is false. If *AtonkB* is counted as true under those conditions, then the second rule for *tonk* will lead from true premises to a false conclusion when *A* is true and *B* is false. But if *AtonkB* is counted as false under those same conditions, the first rule will then lead from truth to falsity.

It is at this point that the warranting holist is likely to intervene in the discussion, and advertise the virtues of his position. After all, he may say, if the defect of these spurious connectives is the impossibility of giving a coherent account of their contribution to semantic value, then what is wrong with them is not obviously anything to do with holism. So why should there not be a warranting holism which, unlike warrant-free holism, insists that any specification of a genuine meaning (perhaps by means of an understanding-condition) must admit a corresponding account of its contribution to semantic value, but which remains a form of holism nonetheless?

Warranting holism faces two closely related problems, which I call the *overdetermination* problem and the *overdiscrimination* problem. We can illustrate the overdetermination problem for the simple case of "and." It is plausible that the understanding-condition for this expression of English involves some kind of mastery of the introduction and elimination rules for conjunction, and that (for a realist, at least) the semantic value of "and" is that classical truth-function which makes those rules always truth-preserving. But for a thinker

who possesses the concept of probability, "and" will feature in other principles too. Such a thinker will have some form of mastery of the principle that if A and B are independent propositions, then $\text{prob}(A\&B) = \text{prob}(A) \cdot \text{prob}(B)$. If we ask what semantic value for "&" would make this principle always correct, again the classical truth-function for conjunction is the answer. So the total set of principles essentially involving "&," acceptance of which is required if the thinker is to understand the expressions they contain, overdetermines the required semantic value.

Why is this a problem? It is a problem for the warranting holist, because it is a datum which can easily be exploited by an opponent of global holism about meaning. The opponent will say that this state of affairs is symptomatic of the fact that only a subset of principles essentially containing "&" are constitutive of its meaning (the introduction and elimination rules). It is they which fix its semantic value, which is then drawn upon and presupposed by someone who starts using the concept of probability in combination with logical connectives like conjunction. This is why the principles of probability can be justified without appealing to those principles themselves as partially fixing the semantic values of the logical constants, the anti-holist will say. Our warranting holist may reply that all principles essentially containing "&" are on a par. But that reply is quite implausible about "&."

The overdetermination problem is on one side of a coin which has the overdiscrimination problem on its other side. The more principles we include as individuative of the meaning of an expression in a person's language, and thereby as contributing to the determination of its semantic value, the wider is the range of cases in which we are precluded from identifying its meaning with that of an expression in the language of a person with either a much richer or a much narrower vocabulary. In one of his earlier writings on the topic, Dummett ascribes to the holist the view that "deduction is useful, because by means of it we can arrive at conclusions, even conclusions of the simplest logical form, which we could not arrive at otherwise" (Dummett, 1978, p. 303). But in fact that is precisely what we cannot do on the holist's view of meaning. According to the holist, when we enrich our vocabulary the meaning of all our expressions changes, and a conclusion we reach with the methods involving a new concept, though it may be grammatically the same as one formulable in the old vocabulary, does not actually have the same meaning as its orthographically identical predecessor. As a result of this overdiscrimination of meanings, the territory in which a global holist can also justifiably, by his own lights, apply a notion of warrant that is unavailable on the warrant-free conception is really quite limited.

6 Local Holisms and Their Source

Sometimes, perhaps always, a thing (property, relation) is individuated in part by its relations to other things, properties, or relations. What it is to be that thing, property, or relation cannot be properly explained without mentioning those other things, properties, or relations. I mention three kinds of case, each familiar from discussions in different areas. First, what it is to be a particular place cannot be explained without mentioning the network of spatial relations in which the place stands. A second plausible example involves mass and force. What it is for something to be the physical magnitude of mass cannot be elucidated without alluding to the fact that things with mass require the action of a force for a change in their motion, and are capable of exerting forces when their state of motion changes. Conversely, the physical magnitude of force is individuated in part by its connections with

the physical magnitude of mass. Third, the property of having a certain linguistic meaning is individuated in part by its connections with the property of believing something with the same content. It is partially constitutive of meaning that it can be used to express beliefs (and even, I would argue, knowledge).[5]

All these claims about what is constitutive or individuative of an object, property, or relation are, in the first instance, claims about things at the level of reference, rather than at the level of sense. They are claims about the things or properties themselves, rather than about those things or properties as thought of in a certain way. It is true that one often formulates such points by saying, for example, "The concept of mass is the concept of a property which, when instantiated requires the action of a force for …." But when one uses such a formulation, one is employing that concept or way of thinking of mass which makes explicit the way in which the property to which it refers is individuated. I say that the claims are "in the first instance" about things at the level of reference, because the point I wish to emphasize does not preclude the possibility that further philosophical analysis may reveal that these constitutive facts about the level of reference may ultimately themselves have an explanation which involves other facts about the level of sense. A parallel may help to make the point. Kripke's arguments have made a strong case that it is necessary, of Peter Serkin, that his father is Rudolf Serkin (Kripke, 1980). This is a *de re* claim about Peter Serkin himself, and not about some mode of presentation of him, and it is arguably constitutive. In any case, let us suppose that it is so, for the sake of the illustration. It does not follow that the ultimate source of *de re* necessities of origin is not some *a priori* principle stating that continuant objects necessarily have their actual origins, a principle whose *a priori* status traces back in part to the sense of "continuant object." That is a separate question, and neither a positive nor a negative answer to it undermines the fact that the original essentialist claim about Peter Serkin involves the man himself, and not some concept of him.

In our three examples, there will be a local holism for the meanings of expressions for (and equally for concepts of) places and spatial relations; for expressions for mass and force; and for expressions for meaning and belief. In each case, what it is to understand an expression for one of these things will have to be given simultaneously with an account of what is involved in understanding (or at least possessing a concept of) the other. These local holisms are entailed by the conjunction of two plausible claims. The first claim is that when one thing (property, relation) rather than another is the reference of a word, there must be some fact about the use of that word which, possibly together with the way the world is, fixes it and nothing else as the word's reference. We can call such a fact about the use of the word or concept "the reference-fixing fact." The second claim is that the reference-fixing facts for words referring to the things and properties in our three examples involve the language-user's rudimentary grasp of the constitutive relations of the thing or property in question to other things or properties. If these claims are sound, then local holisms of meaning are derivative from holisms at the level of the individuation of properties and things.

Grasp of these interrelations is not, of course, and could not be, the only element in an account of mastery of these expressions. There are practical components, involved in the explanation of spatial actions, in the understanding of expressions for places and spatial relations. The same holds for mass and force. Some would argue that there are first-person elements in the grasp of the concepts of meaning and belief. But in all three cases, it is plausible that understanding expressions for the things or properties comprising the local holism involves some form of knowledge of their role in theories capable of explaining respectively, in the three examples, spatial facts, mechanical facts, and facts about intentional action.

* * *

Even if the general arguments for meaning holism are not convincing, there remain a great many intriguing questions about more local holisms. Some of these questions are questions about particular examples. What, for instance, is the relation between practical spatial abilities and mastery of concepts of places and spatial relations? Can a family of practical abilities also display a form of holism? Beyond the questions about particular examples, there are also general questions about kinds of local holism, to which the answers about particular examples are pertinent. Is it always possible in principle to specify the nature of the connections a thinker has to grasp between the properties and objects involved in a local holism without presupposing some mastery of the very concepts involved in the holism? If not, that would seem to count in favor of a version of anti-reductionism about mastery of those concepts. And whatever the answer to that general question, what could or should be the shape of a computational psychology which explains, at the sub-personal level, mastery of a local holism? As far as I know, all these questions are open. Even if global holism is false, the topic of holism deserves to be with us for some time to come.

Notes

1 The arguments for meaning holism have been subjected to a lively critique in Fodor and LePore (1992). Though Fodor, LePore, and I agree in our major conclusions, the arguments for them are rather different, and also venture into different territories. For those wanting a critical overview of discussions about meaning holism, I believe that Fodor and LePore's book and the present chapter will be found complementary, rather than intersubstitutable.

2 Moulines (1986) and Vuillemin (1986) each provide an interesting discussion of the first of these differences between Duhem and Quine, but both pass over the second difference.

3 For an important discussion of the sense of natural kind terms pertinent to these points, see Wiggins (1980, pp. 78–84).

4 It should be noted that my rather strict use of 'analytic' diverges from that of Fodor and LePore (1992). As far as I can see, many of the claims Fodor and LePore make about the analytic/synthetic distinction are ones I would formulate as claims about the *a priori*/*a posteriori* distinction (and would also then accept). This point applies in particular to their discussion of the view that there is no principled distinction between the propositions a person has to believe in order to believe a given content and those he does not (Fodor and LePore, 1992, pp. 24 ff.). It is plausible that rejection of that view does involve some commitment to the existence and applicability of an *a priori*/*a posteriori* distinction. It is not at all so clear that it involves commitment to the applicability of the notion of truth-purely-in-virtue-of-meaning.

5 If meaning and belief do indeed form a local holism, those who have argued from interpretational considerations to meaning holism have mistakenly taken an admittedly important local holism for a global holism.

References

Block, N. 1986. "Advertisement for a semantics for psychology." In *Midwest Studies in Philosophy*, vol. 10, *Studies in the Philosophy of Mind*, edited by P. French, T. Uehling, and H. Wettstein. Minneapolis: University of Minnesota Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davies, M. 1987. "Tacit knowledge and semantic theory: can a five per cent difference matter?" *Mind*, 96(384): 441–462.

Duhem, P. 1962. *The Aim and Structure of Physical Theory*. Translated by Prince Louis de Broglie. New York: Athenaeum.

Dummett, M. 1975. "What is a theory of meaning? (I)." In *Mind & Language*, edited by S. Guttenplan, pp. 97–138. Oxford: Clarendon Press.

Dummett, M. 1978. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Evans, G. 1981. "Semantic theory and tacit knowledge." In *Wittgenstein: To Follow a Rule*, edited by S. Holtzman and C. Leich, pp. 118–140. London: Routledge.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Field, H. 1977. "Logic, meaning, and conceptual role." *Journal of Philosophy*, 74(7): 347–375.

Field, H. 1989. *Realism, Mathematics and Modality*. Oxford: Blackwell.

Fodor, J., and E. LePore. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.

Grandy, R. 1973. "Reference, meaning, and belief." *Journal of Philosophy*, 70(14): 439–452.

Harman, G. 1982. "Conceptual role semantics." *Notre Dame Journal of Formal Logic*, 23(2): 242–256.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.

McDowell, J. 1986. "Functionalism and anomalous monism." In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, edited by E. LePore and B. McLaughlin, pp. 387–398. Oxford: Blackwell.

McGinn, C. 1977. "Charity, interpretation, and belief." *Journal of Philosophy*, 74(9): 521–535.

Moulines, C. 1986. "The ways of holism." *Noûs*, 20(3): 313–332.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. 1993a. "How are a priori truths possible?" *European Journal of Philosophy*, 1(2): 175–199.

Peacocke, C. 1993b. "Proof and truth." In *Reality: Representation and Projection*, edited by J. Haldane and C. Wright, pp. 165–190. New York: Oxford University Press.

Prior, A. 1960. "The runabout inference-ticket." *Analysis*, 21(2): 38–39.

Putnam, H. 1986. "Meaning holism." In *The Philosophy of W. V. Quine*, edited by L. Hahn and P. Schilpp, pp. 405–426. La Salle, IL: Open Court.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1961. "Two dogmas of empiricism." In *From a Logical Point of View*, Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1974. *The Roots of Reference*. La Salle, IL: Open Court.

Quine, W. V. O. 1986. "Reply to Hilary Putnam." In *The Philosophy of W. V. Quine*, edited by P. Schilpp, pp. 427–431. La Salle, IL: Open Court.

Sellars, W. 1974. "Meaning as functional classification." *Synthese*, 27(3–4): 417–437.

Vuillemin, J. 1986. "On Duhem's and Quine's theses." In *The Philosophy of W. V. Quine*, edited by P. Schilpp, pp. 595–618. La Salle, IL: Open Court.

Wiggins, D. 1980. *Sameness and Substance*. Oxford: Blackwell.

16

Metaphor

RICHARD MORAN

Metaphor enters contemporary philosophical discussion from a variety of directions. Aside from its obvious importance in poetics, rhetoric, and aesthetics, it also figures in such fields as philosophy of mind (as in the question of the metaphorical status of ordinary mental concepts), philosophy of science (as in the comparison of metaphors and explanatory models), in epistemology (as in analogical reasoning), and in cognitive studies (as in the theory of concept-formation). This chapter will concentrate on issues metaphor raises for the philosophy of language, with the understanding that the issues in these various fields cannot be wholly isolated from each other. Metaphor is an issue for the philosophy of language not only for its own sake, as a linguistic phenomenon deserving of analysis and interpretation, but also for the light it sheds on non-figurative language, the domain of the literal which is the normal preoccupation of the philosopher of language. A poor reason for this preoccupation would be the assumption that purely literal language is what most language use consists in, with metaphor and the like sharing the relative infrequency and marginal status of songs or riddles. This would not be a good reason, not only because mere frequency is not a good guide to theoretical importance, but also because it is doubtful that the assumption is even true. In recent years, writers with very different concerns have pointed out that figurative language of one sort or another is a staple of the most common as well as the most specialized speech, as the brief list of directions of interest leading to metaphor would suggest. A better reason for the philosopher's concentration on the case of literal language would be the idea that the literal does occupy some privileged theoretical place in the understanding of language generally, because the comprehension of figurative language is itself dependent in specific ways on the literal understanding of the words used. This is at least a defensible claim and, if true, we might then hope for an understanding of figurative language from a theory of literal meaning, combined with an account of the ways in which the figurative both depends on and deviates from it.

The light such an investigation may shed on non-figurative language will derive from the issues which even this mere sketch of their relation raises for the philosophy of language. We will want to know, for instance, about the specific nature of the dependence of the figurative on the literal; and how the comprehension of figurative language is related to, and

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

different from, the understanding of the literal meanings of the words involved. If the theory of meaning in language is, at the least, closely allied with the theory of what understanding such things as sentences consists in, then a question raised by metaphor is how understanding as applied to metaphorical speech is related to understanding in this semantic sense, and whether the same kind of knowledge, such as whatever it is that 'knowing a language' consists in (see Chapter 12, TACIT KNOWLEDGE), applies in similar ways in the two cases. We will want to consider reasons for and against speaking of a difference in *meaning* in connection with metaphor, and whether such distinctive meaning is to be sought for on the level of the word, sentence, or utterance; on the level of semantics, pragmatics (see Chapter 6, PRAGMATICS), or somewhere else.

1 Figurative and Non-figurative: Metaphor, Idiom, and Ambiguity

The familiar subject–predicate form ('X is a wolf, the sun, a vulture …') comprises but one class of metaphors, and neglects various other grammatical forms (such as 'rosy-fingered dawn' or 'plowing through the discussion'), not to mention metaphoric contexts which don't involve assertion at all. And, in general, short, handy examples will not help much in the understanding of, say, literary metaphors whose networks of implications are not discernible outside the verbal environment of a particular text or genre. Nonetheless, even such simple cases can help us to make some provisional distinctions between metaphor and other figurative and non-figurative language. For instance, idioms, such as 'to kick the bucket' or 'to butter someone up,' resemble metaphors in calling for a special reading. If one understands such expressions correctly, one will not expect reference to have been made to any actual bucket or real butter. In a word, they are to be taken figuratively and not literally. (This is so even though, for instance, there is hardly anything wildly paradoxical in the idea of someone kicking a genuine bucket.) But although they both involve giving figurative readings to an utterance, there are important differences in how one comprehends the meaning of an idiomatic expression and the meaning of a metaphor. If you don't know what 'vulture' means or what plowing is, you won't be able to interpret their metaphorical expressions at all. And what one does when one interprets the metaphor is employ what one knows about vultures and what is believed about them to determine what the utterance means on this occasion. This is part of what is meant by the previous suggestion that the comprehension of figurative language is dependent on the literal understanding of the words used.

If idiom is to count as a case of figurative language (which it seems it should, since we can distinguish what it is literally to kick the bucket and the very different thing usually meant by the expression), then this claim of dependence on the literal will have to be amended. For an understanding of the literal meanings of the words that make up an idiom is of very limited usefulness in understanding what is meant, and is sometimes even positively detrimental to such understanding. Someone unfamiliar with the expression will not get very far by employing his understanding of what is known or believed about such things as buckets to figure out what the expression means. And further, if she does know a great deal about the literal meaning of a word like 'moot,' for instance, then, other things being equal, this may well render her less rather than more likely to understand what is meant by the (American) idiomatic labeling of something as a 'moot point' – that is, that the point is

of no current practical import and not worth discussing. What this means is that the meaning of an idiomatic expression is not a function of the meanings of the individual words that compose it; unlike metaphors, they are simply taught to us as wholes, rather than being a matter of individual interpretation on an occasion. (For such reasons, it has been said that "an idiom has no semantic structure; rather it is a semantic primitive." Davies, 1982–1983, p. 68. See also Dammann, 1977–1978.) And again, unlike metaphors, their meaning is simply given: there is no 'open-ended' quality to the idiom's meaning, no special suggestiveness, no call for its creative elaboration. There is a simple, stable answer to the question of what 'kick the bucket' means idiomatically, and that is why dictionaries can have special sections in them for idioms, but not for metaphors (see Cavell, 1976/1969).

Finally, the contrast with idiom enables us to distinguish some issues here concerning paraphrase. It is often said that metaphors, or at least poetic, 'live' metaphors, are not subject to paraphrase, and this is often taken to mean that they are not translatable into another language. However, there is one sense in which it is idioms and not metaphors which resist translation into another language. The overall effectiveness of certain literary metaphors will, to be sure, be influenced by certain language-specific phonetic features; but nonetheless, referring metaphorically to someone as, say, 'shoveling food into his mouth' will be possible wherever they have shovels and food, and words for these things. By contrast, translating the words of the idiom 'to kick the bucket' into Spanish or Korean will not be likely to get across your meaning, or any other meaning. The reason for this, again, is the 'semantic primitiveness' of idiomatic phrases. Since an idiom's meaning is not built up from the meaning of its individual words, this meaning will not be conveyed in another language by means of word-by-word translation (see Dammann, 1977–1978). Naturally, this doesn't mean that some perfectly good sense of 'translation' is not appropriate here. If 'kick the bucket' is one way in English of saying that someone died, then there will be perfectly good ways of translating that idea into Spanish or Korean. So resistance to word-by-word translation is not the same as the inability to express the meaning of the idiom in words of another language.

One way in which the issue of the translatability of poetic metaphor is vexed is through confusion about what might be meant by the idea of a word's acquiring a specifically 'metaphorical meaning'; and this idea will be discussed at some length later. But, in addition, there is some lack of clarity about the relation between paraphrase and translation. If all we mean by paraphrase is the ability to say what one means in other words, then it does seem true that there is a difference between idiom and metaphor here. For, as described above, the idiomatic meaning of some expression can be given in other words in a quite straightforward and definite manner. (Many idioms are euphemisms, after all, whose literal equivalents are all too straightforward.) By contrast, the paraphrase of a live metaphor is much less definite, more open-ended, more dependent on context (including the individual speaker), and more open to the creative interpretation and elaboration of the hearer. What should be noted, however, is that these are all features of the paraphrase of metaphor within a language, and do not carry over any immediate implications for the translation of metaphors across languages. Familiar ideas about the 'essential incompleteness' of any prose paraphrase of metaphor should not cloud the issue, for there is no reason in principle why the very same indefiniteness and open-ended character of a metaphor in English should not show up in its version in another language. Translation need have nothing to do with reducing the live metaphor to a prose paraphrase. And if it is argued that even good translation will not capture all and only the connotations and associations of the original metaphor, it

may be replied that to the extent that this is true at all, it will apply to cases of perfectly literal language as well, from '*Gemütlichkeit*' to 'priggish.' To sum up: within a language, the idiomatic meaning of an expression may be completely given by its literal equivalent, whereas the live metaphor is not reducible to its prose paraphrase; and across languages, an idiom cannot be translated word by word, but only as a fused whole; whereas word-by-word translation of a metaphorical expression may, in fortunate circumstances, preserve the same suggestiveness and 'open texture' as the original. In so far as metaphor involves comparison of things and ideas with other things and ideas, it is something less specifically language-bound than is idiom.

In this respect metaphors also differ from puns, homonyms, and ordinary ambiguity in language. A pun in English, like 'heart' and 'hart,' may be metaphorically exploited by a poet, but is only a homophonic accident until it is so exploited. A translation of the play into another language may well display the same metaphoric comparison, but naturally the phonetic motivation for making just this comparison will be lost with the homophony. Sometimes homophonic words are not only pronounced the same but are also spelled the same, and then we have true homonyms, like 'cape' for a body of land and an article of clothing. An inscription such as 'cape' is ambiguous between the two meanings, which need not be etymologically related at all, and once again this ambiguity may be metaphorically exploited. But neither puns nor homonyms are in themselves examples of figurative language. 'Cape' has (at least) two meanings, but they are both perfectly literal ones, and understanding one of the meanings provides no interpretive clue to the other one.

2 Metaphorical Meaning

Even this brief characterization raises deep theoretical issues, in so far as it has appealed to some notion of 'figurative meaning' at various different stages. In metaphor we interpret an utterance as meaning something different from what the words would mean, taken literally. Often we will want to say that a statement which is wildly false when taken literally is quite true when taken figuratively. And from here it is natural to reason in the following way. Truth-values cannot vary unless *truth-conditions* vary, and if the truth-conditions of an utterance are what determine its meaning, then the literal and the metaphorical interpretations of an utterance amount to differences in meaning (see Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S). The words, or the utterance, have one meaning when intended or taken literally, and another when spoken metaphorically. In addition it was argued, in connection with idiom, that a metaphor can be translated into another language while preserving its metaphorical meaning, and in his original (1962) paper Max Black takes this to imply that "to call a sentence an instance of metaphor is to say something about its meaning, not about its orthography, its phonetic pattern, or its grammatical form" (p. 28).

Thus, some of the motivation for talking about 'meaning-shift' in connection with metaphor is clear enough; and it seems equally undeniable that, quite often, everyday metaphorical speech is successful at communicating something different from what the words, on their literal interpretation, would mean. But our brief characterization of metaphor, especially in its contrast with idiom and common ambiguity, already raises some serious questions for this way of talking about metaphor. For it was pointed out that, unlike the cases of 'kick the bucket' or 'cape,' the different reading we give to 'vulture' (when used, say,

to refer to a certain kind of human predator) is directly dependent on our understanding of the literal meanings of the individual words. Unlike an ambiguous word like 'cape,' then, in metaphor the two meanings must be related somehow. When a *token* of 'cape' is reinterpreted as having one meaning rather than another, the meaning assigned to it on the first reading is excluded, and nothing in the first reading (other than one's dawning sense of its inappropriateness) plays a role in bringing one to the second interpretation. In principle, and often enough in practice, the reader could have hit on the correct interpretation the first time, without considering any possible ambiguity, and nothing would have been thereby lost in her comprehension of what was said.

Such cases of ambiguity explain some of the motivation for individuating words according to sameness of meanings, rather than according to sameness of spelling or pronunciation. (Hence, on this view, the two 'capes' count as different words.) For a speaker does not clarify her intentions by saying she employed the same word, 'bank' (encompassing both meanings), on one occasion to refer to part of the river and on another occasion to refer to where she keeps her money. There is no point expressed in using the 'same word' in these different ways; for the two words are hardly more related in meaning than are 'kinder' in English and *Kinder* in German. In neither this case nor the case with 'bank' need the orthographic identity ever have occurred to the speaker in order to use the words correctly and to communicate her meaning fully.

Contrast this with the case of metaphor. If we think of the words of a metaphorical expression as undergoing a 'meaning-shift' of some kind, it will have to involve a difference of meaning very different from that involved in ordinary ambiguity. For when an expression is interpreted metaphorically, the first interpretation (the literal one) is not canceled or removed from consideration. The literal meaning of 'vulture' is not dispensable when we interpret it metaphorically in its application to some friend or relation. The literal meaning must be known to both the speaker and the audience for the metaphorical point of the epithet to be made. It has everything to do with clarifying the speaker's intentions that she chose this word, with its literal meaning applying to a kind of bird, to refer to this other thing which is not a bird; and when we start to figure out the reason why the speaker is using this word with its literal meaning in this context, we have begun to interpret what she is intending to get across metaphorically. Simply characterizing metaphor in terms of a change of meaning fails to capture the role of the original, literal meaning.

But the dependence of the metaphorical on the literal runs deeper than this, and raises further doubts about the appropriateness of the idea of 'meaning-shift' in metaphor; for the description of interpretation given so far might apply just as well to a situation in which a person is speaking in a kind of code, in which someone has to interpret her utterance in such a way that certain words are to be replaced by specific other ones. He might conjecture that 'vulture' is one of these words, and hit upon the right substitution for it. In such a case we might well speak of the word 'vulture' being given a different meaning or application in this context.

The case of metaphor differs from this in several ways. First, and perhaps most obviously, there is nothing corresponding to a code for a live metaphor, and no rules to appeal to for going from the literal to the metaphorical meaning. Further, in the case of genuine codes the original meaning of the words will normally be incidental, at best, to the new meaning; and in fact, a coined expression with no previous meaning in the language may do just as well, if not better. In metaphor, on the other hand, if we are to speak of a new meaning, this meaning will be something reachable only through comprehension of the

previously established, literal meanings of the particular words that make it up. And this dependence of the metaphorical on the literal is rather special, in ways that exacerbate difficulties with the view of metaphor as involving a change of meaning. For the first (literal) reading of the expression does not simply provide clues to help you get to the second one, like a ladder that is later kicked away, but rather it remains somehow 'active' in the new metaphorical interpretation. It is not similar to a case in which we first got the meaning wrong and have now successfully disambiguated it. Rather, the literal meaning of 'vulture' remains an essential part of the meaning of the metaphorical expression; otherwise one will have no sense of what metaphorical comparison is intended. If something like 'meaning-shift' is involved in this, then we must explain how the literal meaning of 'vulture' could play any role at all in the generation and comprehension of the metaphorical meaning, if it is this very same original meaning that is supposed to have changed (or, to speak a bit less confusingly, if the word has now taken on a different meaning).

It might be thought that we could avoid this problem by referring to an expansion rather than a change of meaning. That way we could retain and rely on the original meaning of the words, and still describe what is going on in terms of some change of meaning. So, for instance, 'vulture' still refers to the same birds it always did, but now, in addition, it also refers to a certain kind of person. The problem with this idea is that while it describes a certain process of linguistic change, it simply isn't what is meant by live metaphor. Words commonly expand and contract in application over time, and this process can take many forms, some of which may indeed involve metaphor at some stage. But the process itself is not inherently metaphorical, and it can proceed for any number of reasons. In earlier times, the word 'engine' applied more narrowly to instruments of war and torture, and not generally to any mechanism that converts energy into force or motion. This expansion in application does not make the latter, contemporary use metaphorical, even if we think that, for instance, certain relations of perceived similarity played a role in the expansion. And in any case, what any such analysis of 'meaning-change' in terms of merely extended application leaves out of consideration is the point insisted on above, the special dependence of the metaphorical on the literal which makes the literal meaning of a word such as 'vulture' still 'active' in the comprehension of its metaphorical use. We are still in need of an account of this 'activity,' to be sure, but there is certainly an essential functional role for the awareness of the literal meaning of 'vulture' in the comprehension of its metaphorical use which has no parallel in the understanding of various other predicates with extended applications. So we still lack an explanation of what could be meant in speaking of 'change of meaning' in connection with metaphor.

These questions will require answers just as much on an account that appeals to speaker-meaning rather than semantic meaning (Searle, 1979; Black, 1979) as it will also on 'extensionalist' accounts, which eschew talk of 'meanings' altogether in favor of reference to different applications of labels (see Goodman, 1968; Elgin, 1983; Scheffler, 1979).

3 Davidson and the Case against Metaphorical Meaning

How might we characterize the dependence of the metaphorical on the literal, specifically the way in which the literal meaning is still 'alive' in the metaphorical application, and avoid making reference to a new metaphorical meaning? And, on the other hand, if we do avoid all such reference, how can we account for the difference in truth-value between the

utterance taken literally and taken metaphorically? Further, if we drop all reference to meaning, then it will be quite unclear how we can make sense of the idea that we correctly understand the speaker as saying (or meaning) something different from what her words literally mean, or that we see metaphor as a vehicle of communication at all.

In a paper that has attracted a great deal of commentary, Donald Davidson has taken this step, and has argued that we should indeed cease talking about figurative meaning in connection with metaphor altogether; and he seems prepared to accept the consequences that follow from this rejection. Early on, he states the thesis of the paper as the claim that "metaphors mean what the words, in their most literal interpretation, mean, and nothing more" (Davidson, 1979, p. 246). He does not mean to deny that metaphor accomplishes many of the same things that philosophers and literary critics have claimed for metaphor (such as the special suggestive power of poetic metaphor, or its capacity to produce insight of a sort that may not be capturable in plain prose), but he denies that these accomplishments have anything to do with content or meaning of a non-literal sort. It will be useful to look more closely at Davidson's paper, for it is an especially forthright and radical response to many of the same problems in accounting for 'metaphorical meaning' that have emerged elsewhere in recent literature on the subject. At the same time we can gain a better appreciation of the costs as well as the benefits of rejecting 'metaphorical meaning.' (Davidson's paper is discussed in Cooper, 1986; Davies, 1982–1983; Fogelin, 1988; Moran, 1989; and Stern, 1991, and there are responses by Black and Goodman in Sacks, 1979.)

The argumentative structure of the paper is not always easy to interpret, but Davidson gives a number of reasons for the denial of metaphorical meaning, some of which are related to the argument given above and which contrast metaphor with common ambiguity. He further argues that positing metaphorical meanings does nothing to explain how metaphors function in speech. If, as he says, a metaphor makes us attend to certain covert features of resemblance (Davidson, 1979, p. 247), it tells us nothing about how this is accomplished to claim that the words involved have some figurative meaning in addition to the literal one. It is not only more accurate simply to say that a fresh metaphor typically produces such effects (in whatever causal manner anything else might do so), but it is also more economical, for we are thereby spared the need to account for what these special meanings are and where they come from. In an ordinary, literal context, appeal to meaning can be genuinely explanatory because there we can have a firm grip on the distinction between what the words mean in the language and what they may be used to do on a particular occasion (to lie, for example, or to encourage, or to complain). However, if we think of what metaphorical language is used for (such as to make us appreciate some incongruous similarity) as itself being a kind of 'meaning,' we lose any sense of this distinction. And yet one of the theoretical virtues of appeal to semantic meaning in the first place is that it enables us to explain something of how these words, with this established meaning and in this context, can be used to perform this particular function on this occasion. That is, a particular established meaning provides both constraints on and possibilities for what a word or phrase may be used to do, and for this reason appeal to such meaning (once it is determined by a given context) can be genuinely explanatory of what the phrase is on this occasion used for. But the only meaning which is distinct and independent of the use on this occasion, and which could play any such explanatory role, is the literal meaning of the phrase. (Various writers have criticized Davidson's argument for assuming a concept of literal meaning that is utterly independent of context, but it seems clear that this is not his view: see Davidson, 1979, p. 260.)

In addition, Davidson argues, when we think of metaphor in terms of the communication of a specific propositional content, we can only have in mind the most dead of dead metaphors, such as referring to the 'leg' of a table. And these, he suggests, are not properly metaphors at all. If the expression 'figurative meaning' points to anything at all, it indicates some special power of metaphor, some striking quality that may be productive of insight or creative elaboration on the part of the audience. The failure to capture anything about the distinctively figurative functioning of live metaphor Davidson sees as a further defect of the idea discussed earlier, that the meaning or application of a term is 'extended' in a metaphorical context. For if we say that the literal application of an expression such as 'vulture' is extended, we have first of all said something false, or at best misleading: as if, now, both some birds and some people were straightforwardly vultures, the way both vultures and sparrows are straightforwardly birds. And in addition, for our trouble, we have failed thereby to capture anything figurative about the whole process. And then, on the other hand, if we say that the metaphorical application of the term has been extended, then we seem to have got no further in our analysis. For we now owe an explanation of what a metaphorical application is, and specifically, how it differs from any other type of application of a term.

(For a different perspective on what are normally thought of as dead metaphors, see Lakoff and Johnson, 1980.)

4 Paraphrase and Propositional Status

The concentration on live metaphor is bound up with another strand of Davidson's case against metaphorical meaning, but one for which it is difficult to determine the weight he wants to give to the various considerations he brings forward. Whatever makes a poetic metaphor 'live,' it is certainly in large part a function of its power of suggestiveness, the fact that the interpretation of live metaphor is open-ended, indeterminate, and not fixed by rules. As Davidson says at the beginning of his essay, "there are no instructions for devising metaphors; there is no manual for determining what a metaphor 'means' or 'says'; there is no test for metaphor that does not call for taste" (Davidson, 1979, p. 245). The creative indeterminacy of live metaphor is one reason why live and dead metaphors differ with respect to the possibilities for paraphrase, or for specifying the meaning in other words. We can fully state what is meant by the 'shoulder' of a road, precisely to the extent that there isn't anything figurative left to the expression. With genuine, or poetic, metaphor the case is quite different, and at various points Davidson seems to be asking, 'How could the sort of open-ended, non-rule-governed character of live metaphor possibly apply to anything legitimately called a meaning?' When we encounter difficulties in applying paraphrase to live metaphor, the reason for this is simply that "there is nothing there to paraphrase" (p. 246). If there were anything said or asserted in the metaphorical expression beyond what it literally states, then it would be just the sort of thing that does submit to paraphrase. As it is, however, what it provides us with beyond the literal is not anything propositional at all.

> It should make us suspect the theory that it is so hard to decide, even in the case of the simplest metaphors, exactly what the content is supposed to be. The reason it is often so hard to decide is, I think, that we imagine there is a content to be captured when all the while we are in fact focusing on what the metaphor makes us notice. If what the metaphor makes us notice were

> finite in scope and propositional in nature, this would not in itself make trouble; we would simply project the content the metaphor brought to mind on to the metaphor. But in fact there is no limit to what a metaphor calls to our attention, and much of what we are caused to notice is not propositional in character. (Davidson, 1979, pp. 262–263)

In this passage, however, Davidson seems to allow that reference to a kind of meaning distinct from the literal would be legitimate if what the utterance got across were "finite in scope and propositional in nature." Then, presumably, we could get a handle on paraphrase, and we could start talking about what was said and what was meant. It was said earlier that it is difficult to settle how much Davidson wants to rest on these considerations; and the reason for this is that, although they run through the entire paper, he also freely admits that it may just as well be said of literal language that its interpretation is not determined by rules (Davidson, 1979, p. 245), and that what it gets across to the audience is often not "finite in scope" (p. 263, n. 17). And, certainly, no theorist wants to deny meaning or cognitive content there. (As far as putting into other words goes, we might also ask how one would paraphrase many perfectly literal statements, such as 'The sky is blue' or 'I can hear you now.') Nor should simple vagueness or indeterminacy in interpretation be thought of as crucial to the issue of meaning, for vagueness itself can be something fixed by the dictionary meaning of a term. For instance, 'house' is a word with a perfectly straightforward meaning, but which allows for a zone of indeterminacy as to just which structures shall count as houses (for discussion of different conceptions of vagueness, see Chapter 28, SORITES.)

If there is to be a genuine case against metaphorical meaning along these lines, then, it seems that we should see the crux of the issue not as concerning indefiniteness as such, but as concerning the question of whether we may speak of propositional content in connection with metaphor. It is certainly true, as Davidson says, that "much of what we are caused to notice is not propositional in character"; but it does not follow from this that the figurative process does not communicate anything that is propositional as well. It seems clear that part of what traditionally raises philosophers' suspicions about the propositional content of poetic metaphor is not the assumption of an incompatibility of content with indeterminacy, but rather the connection of this aspect of the figurative dimension of metaphor with ideas of ineffability, or the essential inability to capture this dimension in words other than those of the specific metaphor itself. When a content or a thought is held to be ineffable, and not simply indeterminate, it is felt that, although one may have a perfectly definite content in mind, it cannot be fully expressed in words. (In fact, in various contexts the sense of indescribability is a response to the highly determinate character, the utter specificity, of what one has in mind.) Or, as in the case of certain poetic metaphors, it may be felt that the idea may be verbally expressed, but only in these very words; or only indirectly expressed, or incompletely hinted at. This sense is certainly something different from simple vagueness, and does raise different questions for the idea that what live metaphor does is communicate some special propositional content. If we agree with Davidson that this problem removes any justification for looking for propositions expressed by metaphorical utterances, then we may still say all we like about the various non-cognitive effects of such utterances, but we will no longer be able to describe metaphor in terms of communication, meaning, or content.

However, ineffability of the sort under consideration here concerns a claim about the specifically linguistic representation of a thought, and does not immediately place something outside the bounds of the propositional unless we have already agreed that a proposition is something essentially linguistic or sentential. Only then will it seem obvious that accepting

an equivalent prose paraphrase is necessary for any part of the metaphor to count as a propositional content. Davidson could be correct when he says, "A picture is not worth a thousand words, or any other number. Words are the wrong currency to exchange for a picture" (Davidson, 1979, p. 263), but it wouldn't follow from this that a picture cannot itself be a representation of a propositional content. For on one standard view of what propositions are, they are "functions from possible worlds into truth values" (Stalnaker, 1972); and on such an account – whether or not it takes reference to 'possible worlds' at face value (see Chapter 31, MODALITY, §3) – pictures, maps, memories, or anything else that represents the world as being a particular way can qualify as propositional representations. (We may thus, in Stalnaker's words, "abstract the study of propositions from the study of language.") If one takes this wider view of what a proposition is, there may be less resistance to considering the possibility of someone with a particular cognitive content in mind, but who is either unwilling or unable to accept an equivalent of it in prosaic language. We could accept Davidson's point about translation into another representational medium, without accepting the identification of the propositional with the sentential.

In fact, for purposes of this discussion, there would be little to complain of in the restriction of propositional content to the meaning of sentences, so long as we kept in mind the various different ways in which the content of a sentence may be indicated and determined in a context, including making essential reference to something extra-linguistic. We may note that many belief reports are only partially verbal reports, with the essential content of the belief being indicated in some other way:

> Many of our beliefs have the form: 'The color of her hair is –,' or 'The song he was singing went –,' where the blanks are filled with images, sensory impressions, or what have you, but certainly not words. If we cannot even say it with words but have to paint it or sing it, we certainly cannot believe it with words. (Kaplan, 1971, p. 142)

Thus, to bring us a little closer to the case of metaphor, a sentence like 'He said it in this voice just like Akim Tamiroff' is in perfectly good order, and expresses a genuine thought. But, of course, it will not communicate much to someone who has never heard of Akim Tamiroff. This particular person and the experience of his voice are essential to the content of the proposition. To someone who has never heard this voice, the speaker may quite straightforwardly be unable to communicate what she means. And it is all too easy to imagine being unable to provide any descriptive equivalent, and that no substitute expression will capture what you want to say. Yet it would certainly be wrong to conclude from this that the speaker has not said or meant anything. (For a defense of the idea of metaphorical meaning, which makes extensive use of the comparison with demonstratives, see Stern, 1985 and 1991. See also Chapter 38, THE SEMANTICS OF PRAGMATICS AND INDEXICALS.)

Similarly, with a metaphorical expression like the well-worn example of Juliet and the sun, reference to the sun is essential to the determination of the content of what Romeo has in mind, and his reluctance to accept any prose paraphrase as capturing all that he means is not itself any reason to deny that he does have something in mind which he is seeking to express in words. Nor would it be right to say that although he does have some content in mind (since we reject the simple sentential view of cognitive content), there must be some confusion involved in trying to express it verbally. Hence, to qualify a concession made earlier for the sake of argument, words may sometimes be the wrong medium of exchange for a picture, but it depends on what we are expecting the words to do. We may not be

entirely satisfied with any descriptive translation of what was said in either the Akim Tamiroff or the Juliet case, but even so it won't follow that "the attempt to give literal expression to the content of the metaphor is simply misguided" (Davidson, 1979, p. 263). As with any attempt to put one's thoughts and feelings into words, it may matter a great deal to try to go as far as one can in this direction. If we can't make sense of this kind of effort at descriptive and expressive fidelity, then we can't make sense of the kind of struggle that goes into the composition of poetic metaphor in the first place, let alone more everyday efforts to put the non-verbal world of experience into words.

(These considerations relate to the debate since Aristotle over whether metaphor and simile are essentially different figures. Fogelin, 1988, and Dammann, 1977–1978, both defend a 'comparativist' view of metaphor, and insist on the distinction between figurative and non-figurative comparisons.)

5 Metaphor and Communication

The discussion thus far has suggested that neither vagueness nor the indeterminacy of the interpretation of metaphor provide good reasons for denying that metaphor has a cognitive content beyond the literal. And, further, even if the difficulties or inadequacies of paraphrase are attributed to a degree of 'ineffability' (and not just indeterminacy) in what is seeking expression, this need not mean that we are not dealing with a genuine propositional content. Naturally, these considerations do not by themselves constitute an account of figurative meaning. Many difficulties remain in making sense of meaning and content as applied to metaphor, and these include various problems that were left hanging in the earlier discussion. For instance, we still need to describe a sense of 'meaning' as applied to metaphor which doesn't reduce to ordinary ambiguity or the expansion of application of a term. We have not yet explained the special dependence of the figurative meaning on the literal meaning, a dependence that has so far only been described metaphorically as the literal meaning's still remaining 'alive' in the figurative context (that is, unlike a code). And very little has been said so far to relate the sense of 'meaning' at stake here to more familiar uses of the term in ordinary speech and in more formal uses in the philosophy of language.

But lest we lose heart at the prospect of these and other problems for explicating the sense of figurative 'meaning,' it would be worthwhile to remind ourselves of how serious the consequences would be of endorsing a fully non-cognitive account of metaphor of the sort Davidson and others have recommended. (The most comprehensive defense of the rejection of metaphorical meaning is David Cooper's 1986 book, *Metaphor*, especially ch. 2.) It is important to Davidson's view that it be seen not as 'no more than an insistence on restraint in using the word "meaning"', but rather as a rejection of the idea that "associated with a metaphor is a definite cognitive content that its author wishes to convey and that the interpreter must grasp if he is to get the message" (Davidson, 1979, p. 262). So, to begin with, any such theory is burdened with the same problems as is non-cognitivism elsewhere in philosophy (see Chapter 31, MODALITY, §4, and Chapter 20, REALISM AND ITS OPPOSITIONS, §4). There will be nothing for understanding or misunderstanding a metaphorical utterance to consist in, nothing to the idea of getting it right or getting it wrong when we construe what the 'figurative meaning' might be. Related to this are non-cognitivism's familiar problems with making sense of the apparent facts of agreement and disagreement in the domain in question; for the rejection of any distinctive content to a metaphorical utterance obscures

understanding of what, for instance, the negation or denial of such an utterance can mean, and such a denial will, in the ordinary case, be a denial of the utterance taken figuratively. If there is nothing to the idea of a distinctive figurative content, then there's nothing for the speaker's audience to be agreeing with or dissenting from, except for the statement taken literally, and agreement or disagreement with that statement is not to the point. Further, if the figurative dimension involves no difference in meaning, but instead simply 'nudges us into noticing some resemblance,' then it's hard to say what differences of meaning we can point to between 'Juliet is the sun,' 'Imagine Juliet as the sun,' or even 'Juliet is not (or is no longer) the sun.' All three sentences succeed in linking the two ideas, but they hardly say the same thing. We might compare such problems with the difficulty for moral non-cognitivism in providing an account of the functioning of moral terms in conditional contexts, when some moral predicate is not being asserted, but is used in the context of reasoning and argument. (For more on these and other criticisms of non-cognitivism as applied to metaphor, see Bergmann, 1982; Elgin, 1983; Kittay, 1987; and Tirrell, 1989, as well as the papers mentioned previously in connection with Davidson.)

The cost of the denial of any specifically metaphorical content, then, seems rather steep, and the case for the banishment of metaphor from the realm of meaning to that of 'use' or the brute effects of utterance seems flawed. It is true that there are many things done in speech that do not involve communication and meaning, but are more purely causal effects of utterance (although, of course, communication is causal in its own way too). We are told, for instance, that metaphor gets us to notice things (similarities or incongruities, or whatever). And it is in terms of such particularities of use that Davidson compares metaphor with the use of language to lie, persuade, or complain. However, a few things must be noted about this comparison. First, it is not at all clear that metaphor is a 'use' of language in this sense at all. It would not, for instance, serve as any explanation why someone said what she did simply to say she was speaking metaphorically. Further, lying or complaining can count as "belong[ing] exclusively to the domain of use" (Davidson, 1979, p. 247) rather than meaning, precisely because whether one says 'It's raining out' to lie or to complain does not affect the truth-conditions of the utterance. But, of course, whether the truth-conditions of an utterance may indeed differ on a metaphorical interpretation is just the point at issue, and cannot be begged at this point.

And when we do speak of metaphor as producing various effects, it is important to note in the context of this discussion that it accomplishes these effects in a quite particular manner, one which involves a relationship between a speaker and an audience, and an interconnected network of beliefs about intentions, expectations, and desires; in short, just the sort of situation that Paul Grice and others have argued is what differentiates a situation of meaning and communication from the other various ways in which beliefs may be acquired (see Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING, especially §5). As Davidson notes, plenty of things, like a bump on the head, can get one to notice or appreciate something, even something profound, and we don't think of all such cases as involving anything like meaning or communication. However, metaphorical speech counts as genuinely communicative (of a content beyond the literal) because, among other things, the figurative interpretation of the utterance is guided by assumptions about the beliefs and intentions of the speaker, intentions which, among other things, satisfy the Gricean formula (intending that the intention be recognized by means of this very utterance). And because we are in this way dependent on beliefs about the speaker's beliefs there is a purchase on the ideas of understanding and misunderstanding what was meant, none of which applies when some non-intentional causal phenomenon succeeds in making one appreciate some fact.

The dependence of the hearer on beliefs about the speaker has several layers. To take the utterance as metaphorical in the first place requires assumptions about the beliefs and intentions of the speaker. Then, even the non-assertoric dimensions of the reception of metaphor (framing one thing in terms of another, the clash of images, and so on) are dependent on what we take the relevant dimensions of the comparison or contrast to be. Lacking any idea of the intended salient features of, say, music, food, and love, we would fail to have so much as a non-assertoric comparison or contrast of these elements, let alone a metaphorical assertion. And finally, the interpretation of the utterance involves assumptions about the speaker's beliefs about the various elements, including her beliefs about their salience to the audience, and about what, if any, particular attitude toward these things is expressed by the metaphor. None of these dependencies obtain with respect to all the other various ways in which the phenomena of the world can cause one to be struck by something or other, and that is the primary reason why we speak of communication, understanding, and misunderstanding in the one set of cases and not the other.

6 Pragmatics and Speaker's Meaning

These and other considerations have led many writers on the subject to identify the meaning of a metaphorical utterance with what is called the speaker's meaning, in contrast with the semantic meaning, of the sentence. The latter notion concerns the meaning of a sentence in a given language, and is standardly understood to be a function of either its truth-conditions or its *assertability conditions*, assuming a certain context. Speaker's meaning, by contrast, concerns what a speaker on an occasion may employ a sentence to imply or communicate, a content that may diverge more or less widely from the content assigned to the sentence by the language. Hence, in ironic speech, for example, a speaker may utter the words 'That was a brilliant thing to say,' in order to communicate something quite different from what the sentence-type means in English. (The example of irony shows the usefulness of separating the issue of 'meaning-change' – which patently does not apply to the words of an ironic utterance – from the issues of communication and cognitivity.)

Speaker-meaning will typically be an instance of what Grice has called 'conversational *implicature*.' Very briefly, Grice sees linguistic behavior as guided by a general Cooperative Principle, which divides into various more particular maxims, such as 'Do not say what you believe to be false,' or 'Be relevant,' and which speakers expect to be obeyed in conversational exchange. Naturally, any such maxim may fail to be observed on a given occasion (people do tell lies, for instance). But what is important to Grice's story is the different ways in which a maxim may not be observed. For it may be that it is not followed either through sheer carelessness, or because the speaker is 'opting out' of the conversational exchange altogether, or, most importantly here, the speaker may 'flout' a maxim. In such a case the speaker makes it manifestly clear that, on one level at least, she is intentionally violating some maxim. In the above example of speaking ironically, the speaker takes it to be clear to the audience that she does not think what was just said was brilliant, and yet here she is, uttering a sentence with that very meaning. Hence she is flouting one of Grice's 'Maxims of Quality' ('Do not say what you believe to be false'). At this point it is up to the hearer to construe what the point of the utterance could be, and what other proposition(s) may be intended. The general assumption of the Cooperative Principle is retained, but the hearer now looks for what proposition may be implicated by this utterance. Thus conversational

implicature is a means of communicating something different from the literal, semantic meaning of the sentence uttered.

Taking this general approach, John Searle takes the general formula for metaphor to be: A speaker utters a sentence with (semantic) meaning 'S is P,' but does so in order to convey (or 'implicate') a different proposition, namely 'S is R.' In Searle's example (1979), someone says 'X is a block of ice' in order to convey the very different proposition that X is emotionally unresponsive and so forth. In most cases it will be the manifest or categorical falsity of the sentence taken literally that cues the audience to interpret the utterance as implicating something metaphorically. The main questions for which Searle takes a theory of metaphor to be responsible are, then, how an utterance is recognized as metaphorical (rather than ironic, say), and what principles the hearer employs to compute the speaker's meaning from the meaning of the sentence uttered, combined with the context of utterance.

An account of this general form may, then, offer us a sense of 'meaning' as applied to metaphor, which does not entail that a linguistic entity as such somehow contains within itself a metaphorical as well as a literal (semantic) meaning, but one which is nonetheless a sense of 'meaning' which bears some important relation to meaning in the strictly semantic sense. It also offers some understanding of the special dependence of the figurative on the literal, in that it is only through comprehension of the literal meaning of the statement that the hearer may reach the secondary meaning 'implicated' by the utterance. And the principles that guide this interpretation will involve appeal to features of resemblance, contrast, context, and emotional attitudes toward the subject that make the relation between literal and figurative meaning very much unlike the relation between a word and its substitution in some code.

In addition, such a view need not subsume 'figurative meaning' under the categories of simple ambiguity and ordinary expansion of meaning. Gricean implicature involves there being some point to the speaker's application of this phrase, with this literal meaning, in this context in order to convey something quite different. Common ambiguity (such as, say, homonymy) does not involve any such point or communicative intent. Nor need there be any such point in the case of the ordinary expansion of the application of a term. In some cases there may be some such point to the expanding, but often there will not be. When there is some point to the extension (as in, for example, the extending of 'mouth' to parts of bottles and rivers) the motivation may simply concern some perceived similarity between the various things now referred to by the same term (Davidson's example). In those cases the theorist of 'speaker-meaning' will indeed need to distinguish the point of metaphorical speech from that of ordinary expansion of the application of a term without any communicative point; otherwise he fails to distinguish figurative meaning from some forms of ordinary ambiguity. On the other hand, one may not want to distinguish the two cases too sharply, because metaphor is, after all, one of the vehicles of the normal extending of the application of words. Sometimes when metaphors die, their death involves the alteration of the ordinary dictionary meaning of a term, as in the case of 'mouth.' This phenomenon is, in fact, a further problem for any view that denies any distinct cognitive content to live metaphor. For it is clear that part of the meaning of the word 'mouth' is different now from what it was prior to the development of the metaphor, and yet we would not be able to say where the difference in meaning came from if the metaphor had no content aside from the (old) literal one when the metaphor was alive. By the same token, this would also, of course, oblige the theorist of speaker-meaning, for whom the distinction between it and semantic meaning is crucial, to say something about the diachronic story of how speaker-meaning becomes 'regularized' over time and merges into an altered semantic meaning of the term.

There are thus some promising features of this general approach, but its application and explanatory power also seem to have some significant limitations. First of all, it's not clear, on Searle's version of the theory anyway, that much has been said to elucidate the specifically figurative dimension of metaphor. If what one is doing in speaking metaphorically is saying (or making-as-if-to-say) 'S is P' in order to convey the different proposition 'S is R,' then it is hard to see how anything in the way of special insight or enhanced apprehension of the subject is achieved in this way. And it doesn't seem enough to make up for the flat quality of the analysis to add, as Searle does, that the speaker may intend "an indefinite range of meanings, S is R1, S is R2, etc." (Searle, 1979, p. 115). No degree of indefiniteness alone will add up to power or insightfulness. (And if one is skeptical of the claims made for insight and metaphor, the criticism would remain that even the appearance of power or insight – which surely does require explaining – seems to find no place in this account.) Related to this is the problem, common to many accounts that want to emphasize the cognitive aspect of metaphors and their role in assertion, that the account seems derived from the consideration of only the dead and dying among metaphors. And even this class is normally restricted to examples of the familiar subject–predicate form; whereas, clearly, a major part of the theoretical interest in metaphor concerns the desire to understand what is deeply right or expressive or illuminating in such occurrences of live metaphor as in the dense figurative networks of literature, which need not involve any phrases in subject–predicate form, or be part of any statement of fact (either real or pretended). (By comparison, we might ask here how a caricature or a gesture can be 'right,' expressive, or illuminating.)

A further way in which the 'live' quality of live metaphor seems to escape this analysis is in the account of how interpretation proceeds and what the derived meaning consists in. For if the meaning of a metaphorical utterance is the speaker's meaning, and the latter is a function of the intentions of the speaker in making the utterance, then the meaning of metaphor in general will be confined to the intentions of the speaker. Interpretation of the metaphor, then, will be a matter of the recovery of the intentions of the speaker. This may do well enough for instances of well-worn metaphor with little suggestive power left, but it gives the wrong picture of the interpretation of live metaphor. As Cooper says, in criticism of the 'speaker's meaning' view, "even a quite definite speaker-intention does not finally determine the meaning of a metaphor" (Cooper, 1986, p. 73). It is consistent with this criticism to insist, as claimed earlier here, that the interpreter of a metaphor is dependent on various assumptions about the beliefs and intentions of the speaker, and that this is required even to achieve a sense of what sort of figurative comparison is relevant. For it does not follow from this claim that the interpretation of metaphor is restricted to the recovery of the speaker's intentions. The interpreter may need to presume various things about the beliefs of the speaker for the metaphor to succeed in picturing one thing in terms of another; but once that perspective has been adopted, the interpretation of the light it sheds on its subject may outrun anything the speaker is thought explicitly to have had in mind. And on the other hand, from the point of view of the speaker, the restriction to speaker-meaning seems inadequate, in that it construes metaphor as a kind of shorthand or mnemonic device for a given set of beliefs that she wishes to convey. What such a picture leaves out of consideration is the role of metaphor in thought, the fact that the composition of live metaphor is undertaken in the expectation that it will lead one's thoughts about the subject in a certain direction: that it will be productive of new thought about it, and is not just a convenient summing-up of beliefs one already has.

(The comparison of metaphor with models in science has inspired work on metaphor as a vehicle, and not just a repository, of thought. On this, see various papers in Ortony, 1979. This general point, however, is not restricted to the case where a metaphor functions as a kind of explanatory model, but applies as well to the composition of metaphor in everyday and poetic cases, where it is not functioning as a model for explanation. This aspect of 'metaphorical thought' has received considerably less attention in recent philosophy.)

7 Metaphor, Rhetoric, and Relevance

We have arrived, then, at a familiar point of tension for theories of metaphor. On the one hand, there is the desire (widespread, but not universal) to see metaphor as a cognitive phenomenon and hence as having a describable role in such activities as assertion, communication, and reasoning. But on the other hand, theories of metaphor that seek to defend and define this cognitive role often end up obscuring the very features of metaphor that make it an object of theoretical interest in the first place: its figurative power; the role of metaphor in expressing or producing insight of some kind; or the special open-ended role of the interpretation of live metaphor. It is no surprise, then, that 'non-cognitive' theorists like Davidson emphasize the difference between live and dead metaphors, whereas 'cognitive' theorists often either downplay the distinction or deal with examples of metaphor that might as well be dead.

To make progress from here, it may be useful to reorient our approach to the whole phenomenon, to consider cognition and communication outside the context of strictly linguistic activity, and to begin investigating them from this broader perspective prior to explicit theorizing about the case of metaphor. That is, instead of taking the determinate proposition expressible in a simple sentence as our paradigm, and then asking how closely metaphor may or may not approach this model, we might begin with communicative situations that are non-verbal, indefinite, and unstructured, and ask where we might locate metaphorical speech on a continuum of cases from there to explicit, literal speech. This is more or less the approach taken by Dan Sperber and Deirdre Wilson in their 1986 book, *Relevance*, and in subsequent publications on rhetoric and communication. They see linguistic communication as but one variety of the larger class of what they call 'ostensive–inferential communication,' which encompasses "behavior which makes manifest an intention to make something manifest" (Sperber and Wilson, 1986, p. 49). (Their account has obvious points of contact with Grice's work, as well as important differences, and these are discussed in the book.) The breadth of the category of communication they employ, and the distance from the sentential paradigm, can be seen from one of their first examples of ostension, many of whose features have come up for discussion in the case of metaphor. Two people are newly arrived at the seaside, and one of them opens the window of their room and inhales appreciatively and 'ostensively,' that is, in a manner addressed to the other person. This person thus has his attention drawn to an indefinite host of impressions of such things as the air, the sea, and memories of previous holidays.

> [Although] he is reasonably safe in assuming that she must have intended him to notice at least some of them, he is unlikely to be able to pin her intentions down any further. Is there any reason to assume that her intentions were more specific? Is there a plausible answer, in the form of an explicit linguistic paraphrase, to the question, what does she mean? Could she have achieved the same communicative effect by speaking? Clearly not. (Sperber and Wilson, 1986, pp. 55–56)

If this sort of situation is accepted as an example of communication, we can see how many of the features of metaphor which are thought to stand in the way of any cognitive account find a natural place here; and the case of explicit, literal, verbal communication looks more like the special case. That is, the way may be open to see some types of verbal communication (such as, for example, figurative language) as sharing many of the features of this non-verbal example. Thus we could see metaphorical speech as involving dependence on beliefs about the speaker's intentions, but not restricted in its interpretation to the recovery of those intentions. We could speak of a content that is communicated, but which is to a significant degree indeterminate, resistant to paraphrase, and open to the elaborative interpretation of the hearer. And, since the account does not assume literalness as a norm, we could avoid the implication of a generally Gricean approach that speaking figuratively, for all its utter pervasiveness in everyday speech, must involve transgression of some sort, or the violation of linguistic rules. Or, to quote Sperber and Wilson (1986, p. 200), "[T]here is no connection between conveying an implicature and violating a pragmatic principle or maxim." (See also Cooper, 1986, on the 'perversity' objection to speaker-meaning theories. It should be noted that the rejection of the normative presumption of literalness does not entail the rejection of the previously described dependence of the figurative on the literal, that is, the idea that knowledge of the literal semantic meanings of the words involved is necessary for the composition or comprehension of metaphor.)

Here we can do no more than indicate a few of the main themes of their approach which relate to the case of figurative language. Sperber and Wilson see implicatures as being conveyed in speech not through a presumption of either literality or obedience to conversational maxims, but through the guarantee of relevance which, they claim, any act of ostensive communication carries with it. Such acts will lie on a continuum of cases from communication of an impression to coded information, from showing to saying. In fact, it is internal to this approach that various dimensions of assessment, normally construed categorically, will be such as to admit of differences of degree: literality and figurativeness, evocativeness, susceptibility to paraphrase, and degree-of-intendedness.

Relevance, as defined by them, concerns the value of information gained, in light of the cognitive 'cost' to the hearer of assimilating that information. (As the quoted example indicates, however, 'information' is a suitably broad notion here too.) The guarantee of relevance may go unfulfilled, of course, but it is different from a maxim that one either seeks to conform to or not. Relevance is guaranteed in the sense that any act of ostensive communication involves a claim on the attention of another person, and any such claim itself communicates the presumption that this attention is somehow worth the effort. Implicatures themselves may be weak or strong, far to seek or immediately obvious, and are related to each other in various ways. In this way, we may begin to have at least a useful description of the functioning of live or poetic metaphor, where the effort at interpretation generates a penumbra of stronger and weaker implications which in turn lead to others, more or less remote from the immediate inferential consequences of the utterance, but which are pursued in so far as the presumption of relevance is rewarded. Dead metaphors will be those with a relatively small network of implications, immediately comprehended at small cost. Along these lines, then, we may begin to be able to say a few things about what the figurative power of poetic metaphor consists in, and what claims can be made for it both as productive and as expressive of insight of various kinds (including, for instance, marking the difference between the 'live' or fully-felt appreciation of some fact, and its merely 'intellectual' apprehension).

* * *

A final note. The alternatives to non-cognitivism discussed here have been drawn from theories of conversation, or the pragmatics of language, rather than from semantic theories for natural languages. However, the semantic/pragmatic distinction in philosophy of language is itself a complex matter and a subject of controversy (see Chapter 6, PRAGMATICS). Thus, mention should be made of recent 'cognitive' accounts which explicitly challenge the assignment of figurative meaning exclusively to either one level of analysis or the other. So, for instance, Kittay (1987) describes her 'semantic field' theory of metaphor as one that moves between semantic and pragmatic accounts. And the work of Stern mentioned earlier belongs to a broadly semantic account, but finds the analysis of both metaphor and demonstratives to require "a notion of meaning one level more abstract than truth conditions" (Stern, 1991, p. 40). In these as well as other ways, discussed previously, we can see the theory of figurative language prompting the rethinking of some of the basic concepts in the philosophy of language generally.

References

Bergmann, M. 1982. "Metaphorical assertions." *Philosophical Review*, 91(2): 229–245.

Black, M. 1962. "Metaphor." In *Models and Metaphors: Studies in Language and Philosophy*. Ithaca, NY: Cornell, pp. 25–47.

Black, M. 1979. "How metaphors work." In Sacks, 1979, pp. 181–192.

Cavell, S. 1976. "Aesthetic problems of modern philosophy." In *Must We Mean What We Say?* pp. 73–96. Cambridge: Cambridge University Press; New York: Scribners, 1st edn, 1969.

Cooper, D. 1986. *Metaphor*. Oxford: Blackwell.

Dammann, R. M. J. 1977–1978. "Metaphors and other things." *Proceedings of the Aristotelian Society*, 78: 125–140.

Davidson, D. 1979. "What metaphors mean." In Sacks, 1979, pp. 29–46. Also reprinted in *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 1984, pp. 245–264.

Davies, M. 1982–1983. "Idiom and metaphor." *Proceedings of the Aristotelian Society*, 83: 67–85.

Elgin, C. 1983. *With Reference to Reference*. Indianapolis: Hackett.

Fogelin, R. 1988. *Figuratively Speaking*. New Haven: Yale.

Goodman, N. 1968. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis: Hackett.

Kaplan, D. 1971. "Quantifying in." In *Reference and Modality*, edited by L. Linsky, pp. 112–144. Oxford: Oxford University Press.

Kittay, E. F. 1987. *Metaphor: Its Cognitive Force and Linguistic Structure*. Oxford: Oxford University Press.

Lakoff, G., and M. Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Moran, R. 1989. "Seeing and believing: metaphor, image, and force." *Critical Inquiry*, 16(1): 87–112.

Ortony, A., ed. 1979. *Metaphor and Thought*. Cambridge: Cambridge University Press.

Sacks, S., ed. 1979. *On Metaphor*. Chicago: University of Chicago Press.

Scheffler, I. 1979. *Beyond the Letter*. London: Routledge and Kegan Paul.

Searle, J. R. 1979. "Metaphor." In *Expression and Meaning: Studies in the Theory of Speech Acts*, pp. 76–116. Cambridge: Cambridge University Press.

Sperber, D., and D. Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.

Stalnaker, R. 1972. "Pragmatics." In *Semantics of Natural Language*, 2nd edn, edited by D. Davidson and G. Harman, pp. 380–397. Dordrecht, Netherlands: Reidel.
Stern, J. 1985. "Metaphor as demonstrative." *Journal of Philosophy*, 80(12): 677–710.
Stern, J. 1991. "What metaphors do not mean." In *Midwest Studies in Philosophy*, vol. 16, edited by P. A. French, T. Uehling, Jr, and H. Wettstein, pp. 13–52. South Bend, IN: University of Notre Dame Press.
Tirrell, L. 1989. "Extending: the structure of metaphor." *Noûs*, 23(1): 17–34.

Further Reading

Cohen, L. J., and A. Margalit. 1972. "The role of inductive reasoning in the interpretation of metaphor." In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 722–740. Dordrecht, Netherlands: Reidel.
Cohen, T. 1975. "Figurative speech and figurative acts." *Journal of Philosophy*, 72(19): 669–684.
Grice, H. P. 1975. "Logic and conversation." In *Speech Acts,* vol. 3 *of Syntax and Semantics*, edited by Peter Cole and Jerry L. Morgan, pp. 41–58. New York: Academic Press.
Isenberg, A. 1973. "On defining metaphor." In *Aesthetics and Theory of Criticism: Selected Essays of Arnold Isenberg*, edited by W. Callagan, L. Cauman, and C. Hempel, pp. 105–124. Chicago: University of Chicago Press.
Sperber, D., and D. Wilson. 1985–1986. "Loose talk." *Proceedings of the Aristotelian Society*, 86: 153–171.

Postscript

ANDREW MCGONIGAL

1 Metaphor, Meaning, and Language: Positive Developments

A good theory of metaphor would tell us about the relationship between metaphor and language. In particular, it might help us address the following kinds of foundational questions:

Distinctness: Do metaphors linguistically express representational contents distinct from those that a literal interpretation would assign?
Distinctiveness: Does effective linguistic metaphor typically bring off something which is hard to achieve by literal means?
Dispensability: If we were unable to think and talk metaphorically, would we thereby be significantly cognitively or practically impoverished?
Demands: What kinds of theoretical constraints and explanatory challenges should a satisfying account of linguistic metaphor meet? Which developed account best meets those demands?

There is still no consensus as to how questions like these are best answered. But we are gradually getting clearer on the types of considerations that count for and against affirmative answers in each case. For example, most contemporary theorists of metaphor accept the following explanatory tasks.

a. Address the Error-Theoretic Challenge

Do metaphorical utterances linguistically express metaphorical meanings? Realist semantic theorists say yes: the metaphor *semantically expresses* some distinct metaphorical content. Realist pragmatic theorists say yes: the metaphor maker *pragmatically conveys* some such content. Anti-realist error theorists say no: the metaphorical utterance is at most linguistically related to a literal meaning, if to any. Other representational contents might be relevant to the metaphor – for example, as specifying its point, purpose, or significance – but not any old relation to a propositional content is a *linguistic* relation.

Donald Davidson (1979) gave the simplest and most influential argument for error theory:

(i) we don't need to appeal to metaphorical meanings to explain how linguistic metaphors are made and aptly interpreted, and
(ii) it isn't clear that, as standardly construed, they even *could* adequately explain those practices, so
(iii) we thus have no good theoretical reason to believe in them.

Error theory continues to attract defenders and adherents (see, e.g., Reimer, 2001; Lepore and Stone, 2010; McGonigal, 2008). Even their realist opponents, however, recognize the need to address the detail of Davidson's arguments, and to accommodate the insights of his position.

b. Respect the Active Role of Literal Terms and Context

Davidson stressed two key structural features of metaphor. First, the sense of literal terms plays an active shaping role in securing proper uptake of the metaphor. Standardly, someone who is in a position to fully understand

> A hungry sound came across the breeze,
> So I gave the walls a talking

draws upon their knowledge of the literal semantic contribution of the word "hungry." Second, the interpretation of metaphorical utterance is highly context dependent. When Stevens tells us that poetry is a pheasant disappearing in the brush, or Plath tells us that a pheasant is a little cornucopia, informed appreciation of those metaphorical utterances requires sensitivity to the details of the local context in which they are offered up for interpretation.

Realists about metaphorical meaning have attempted to accommodate these insights. Contextualists point out that standard treatments of many features of literal language – paradigmatically, the semantics of indexicals and demonstratives – make essential appeal to contextual features of token utterances. Different forms of contextualism about metaphor stress appeal to different models of contextual interaction. In one of the most detailed and developed accounts, Joseph Stern draws upon elements of Kaplan's semantic treatment of demonstratives. He argues that communication through linguistic metaphor involves tacit knowledge of a Kaplanian *character*; a rule that determines, for each context C, the content of the metaphor in C. Using *Mthat [E]* to lexically represent the metaphorical

interpretation of some literal expression E, Stern's view is that (i) *Mthat [E]* serves to determine different metaphorical meanings in different contexts (ii) in a manner which is systematically sensitive to the presuppositions of cooperative conversational participants in those contexts. In this way, his theory aims to (iii) account for the active role of the literal meaning of E in helping fix the interpretation of metaphor in context, while also (iv) allowing metaphorical meaning to be highly sensitive to features of context. (Leezenberg, 2001; Recanati, 2004; Carston, 2002; and Bezuidenhout, 2001, appeal to alternative features of context-dependent language.)

Contemporary pragmatic theorists in contrast (including Camp, 2006a; Denham, 2000) locate the active role of literal sense in the meaning of the sentence uttered, rather than the speaker-meaning thereby conveyed. Fine-grained sensitivity to context enters into the story of how that implicit speaker-meaning is recoverable by the interpreter. Quite how this is possible is a pressing issue: see Sperber and Wilson (1995) for criticisms of the standard Gricean account.

c. Integrate the Account with Broader Theoretical Consensus about Linguistic Communication

An influential view of linguistics presents it as

(a) appealing to a small set of sparse, distinctive properties and relations (*X is the semantic value of W*; *Maxim M governs cooperative communication between S and L*) in
(b) offering explanations of projectible, systematic, counterfactual-supporting linguistic regularities (*compositionality*; *systematicity*; *productivity*; *communication*; *presupposition*).

We thus have system and structure on both sides of the explanatory relation. Linguistic theory explains the comparative orderliness of the data in terms that appeal only to a small set of tightly characterized explanatory posits.

Contemporary realists and anti-realists typically agree that context matters to the interpretation of metaphor. They disagree about whether it does so in a way that is orderly enough to fall within the domain of linguistic explanation as described above. For example, error theorists often hold that the link between successful metaphor and its appropriate interpretation isn't systematic enough to be well explained by appeal to semantic or pragmatic rules. On this view, our ability to construct and interpret metaphor effectively doesn't consist in mastery of repeatable, projectible, distinctively linguistic mechanisms, such as Stern's *Mthat* operator, Gricean implicature, or pragmatic enrichment. The success and significance of individual metaphors might be explicable on a case-by-case basis, but not in virtue of our tacitly grasping linguistic rules that determine how an arbitrary metaphorical expression will be paired up with a distinctive, non-literal meaning.

Realists about metaphorical meaning generally concede that this point presents a serious challenge. They agree that the word "pheasant" in Stevens's metaphor does not take on a new stable metaphorical sense, apt for compositional redeployment across a wide range of linguistic contexts. But they stress the extent to which ordinary literal communication relies on mutual recognition of communicative intent, and skillful accommodation to a shared take on the world. These pragmatic abilities might be hard to systematize in a robustly

predictive way, and yet we might feel compelled to accept that some such set of capacities play an ineliminable role in linguistic communication. (See Chapter 6, PRAGMATICS, and Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS.)

This goes some way to answering the general meta-semantic question of how metaphorical meaning is possible at all. Since literal communication is actual, the capacities it requires must be actual: if those are in turn the core capacities drawn upon in metaphorical interpretation, then that's evidence that metaphors could at least in principle express meanings. However, this line of thought doesn't do much to help explain why metaphorical utterances are assigned the *particular* contents that the realist wants to attribute to them. Realists are sensitive to these concerns, and have tried to offer more detailed accounts of the mechanisms that generate interpretations of particular metaphorical utterances. Nevertheless, even the most developed of these typically accept that they do not have a fully satisfying response to the error-theoretic challenge from anti-systematicity.

d. Account for an Accepted, Detailed Range of Linguistic Phenomena

The most effective responses to the error theorist cite a range of linguistic data that is putatively best explained by the existence of metaphorical contents. For example, Stern has emphasized the behavior of metaphor embedded in tensed and modal constructions such as:

> "Greg has always been a failed attempt, a stuffed shirt, a shirt-shaped hole in the air."
> "Yeah, if Rolls Royce started building really terrible cars, he would be the Rolls Royce of people."

and the zeugmatic effect of verb phrase ellipsis in constructions such as

> "Achilles is the sun, and Juliet is too."

in motivating his realist semantic view of metaphorical meaning. Patterning of speech acts such as assertion, rejection, extrapolation, question and answer, provide another potential range of data (see, e.g., Bergmann, 1982). A variety of issues in lexical semantics also seem to cast light on the nature of metaphor. Michael Glanzberg (2007) has suggested that while lexical items such as nouns and verbs can be reinterpreted metaphorically, linguistic items being deployed in 'functional' categories such as quantifier expressions and tenses cannot. If true, then this *prima facie* surprising contrast between, say, "more than" and "larger than" seems to call out for explanation. More generally, Alison Denham (2000) has suggested that pragmatic realism about metaphorical meaning can help explain the role of metaphor in plugging certain lexical gaps.

e. Integrate Account with Psychological Data

The interaction between linguistics, psychology, and philosophy provides another rich seam of data. Peacocke (2009) argues that when we experience one thing metaphorically-as another our experience thereby has a distinctive kind of representational content. If this is correct, then good theory of metaphor should tell us how such metaphorical Thoughts are

related to linguistic utterances. Peacocke's own view is that the connection between metaphor and language is somewhat secondary and indirect:

> The second component of my account is a conception of metaphor as essentially non-linguistic, as something cognitive that can be present in many different types of mental state and event. Metaphor can enter thought; it can enter imagination; and it can enter perception. We can think of life as a journey; we can imagine an atom as consisting of a star and orbiting planets; we can perceive modern windmills in a wind farm as an army of warriors … We have metaphors in language only because we need a device for expressing these mental states whose content involves metaphor. Understanding a metaphor expressed in language involves thinking or imagining whose content is a metaphor. There would be no metaphorical language if there were no mental states whose contents involve metaphor. (Peacocke, 2009, p. 260)

Even if this understates the constructive meta-semantic role played by distinctively linguistic features such as rhyme, rhythm, literary allusion, linguistic echo, and cliché, a completed theory of metaphor surely ought to address the kinds of cognitive phenomena that Peacocke emphasizes.

Evidence from sub-personal psychological processing has also seemed apt to constrain theories of linguistic metaphor. For example, facts about the comparative processing speed of metaphorical and literal language have been thought to count against a pragmatic view (Camp, 2006b, gives a helpful overview). Again, if it is true that some autistic subjects who smoothly understand metaphor struggle to understand irony, but never vice versa, then this would seem to speak against simple Gricean treatments that treated the two cases analogously (Carston and Wearing, 2015).

2 Distinctiveness and Dispensability: For and Against

Elizabeth Camp (2006a) holds that (i) the employment of linguistic metaphor typically involves speakers meaning something different than what they say, (ii) such metaphorical utterances do not do anything distinctively different in kind from literal ones but nevertheless (iii) are essential for grasping certain important propositions. This combination of views can at first seem puzzling: how can the formulation and interpretation of metaphor fail to have distinctive linguistic results, and yet be indispensable for certain tasks? Camp's answer is that what metaphor users mean can be in some sense *expressed* in literal terms, but not *paraphrased*. Paraphrase requires more: provision of a specification of the content of the speaker's illocutionary intention in terms explicit enough that inability to grasp it would indicate some form of semantic incompetence. For Camp, effective selection and interpretation of metaphor is instrumentally essential for our coming to understand the proposition expressed, and so no such literal paraphrase is available.

We might wonder why paraphrase in this sense is what matters. James Grant (2010) has complained that Camp does not establish that metaphor is an essential route to the content of any cognitive or communicative acts. Suppose that Camp is right that

> (M1) When he finally walked out the door, I was left standing on the top of an icy mountain crag, with nothing around me but thin cold air, bare white cliffs, and a blindingly clear blue sky.

cannot be adequately paraphrased by some literal utterance of the form

> (P1) I felt an emotion which was like the way it would feel physically to stand on top of an icy mountain crag, with nothing around me but thin cold air, bare white cliffs, and a blindingly clear blue sky.

because such a specification would still 'rely at least implicitly on the original metaphor' and so fail the explicitness condition on paraphrase:

> if 'like' expresses a substantive relation which holds just in case a particular, contextually salient similarity holds between the two objects … then [P1] implicitly builds those similarities into its content. It may then succeed in capturing the speaker's intended content, but it arguably also fails to be fully explicit, in much the way that 'He's ready' fails to specify its implicit argument. (Camp, 2006a, p. 12)

Even if every such putative paraphrase failed, Grant objects, it is still left open that some of them express the relevant illocutionary content implicitly. And if they do so, then how can the deployment of *metaphor itself* be cognitively or communicatively essential? As Grant notes, communicating a metaphor's content implicitly is different from implicitly relying on a metaphor that communicates the same content. Why should we concern ourselves with unparaphrasability, as long we can systematically pair off every metaphor with a literal statement expressing the same representational content?

Grant presses a related objection against Steven Yablo's claim that some metaphors are 'representationally essential' – that is, that the best way to specify a given metaphorical meaning might involve redeployment of that very metaphor. For example, Yablo (1998) suggests

> It seems at least an open question, for example, whether the clouds we call angry are the ones that are literally F, for any F other than 'such that it would be natural and proper to regard them as angry if one were going to attribute emotions to clouds.'

Even though he thinks that adequately specifying the content of certain metaphors may require redeploying metaphor, Yablo does offer a meta-semantic account of how metaphors derive their content. This is a version of Kendall Walton's view of metaphor as a form of 'prop oriented make-believe' (Walton, 1993). On this view, when we use a metaphor, we represent a situation as having just those properties that would make our utterance appropriate in a certain game of make-believe, one that the metaphorical utterance itself invites or suggests. But Grant seeks to turn this Walton-style explanation against Yablo's claim that metaphorical contents cannot be adequately specified without metaphor. He argues as follows:

(i) Yablo thinks that we need metaphor to access the ensemble of worlds that comprise the metaphorical content of the metaphor S.
(ii) But we can use Yablo's own description of the mechanics of content-generation to pick out the relevant ensemble: namely, the worlds W that legitimate the pretense-worthiness of S.
(iii) That description is not itself metaphorical.
(iv) So *contra* Yablo, we do not need metaphor to access the ensemble of worlds that comprise the metaphorical content of the metaphor S.

This style of objection, however, doesn't seem to give due weight to the fact that the relevant account of paraphrase relates it to linguistic *understanding*. On Camp's view, saying that linguistic metaphors are contentful but unparaphrasable is denying that successful exercise of basic linguistic competence and rationality puts one in a position to understand them, to know what speakers mean by them. Additional psychological capacities are required, and it makes sense to wonder what they are, and how and when they are deployed. That's a different project than aiming to systematically assign suitably related, intensionally equivalent truth-conditions, but one that seems well motivated.

Consider an analogy from the philosophy of mind. Perhaps a judgment with the conceptual content I AM HUNGRY is true if and only if a hungry judger thereby *self-ascribes* the property of hunger. In that sense, the truth-conditions are the same whether we employ the first-person, or an unanalyzed concept of self-ascription. But that wouldn't show that first-person thought was cognitively dispensable. For all that the equivalence shows, being able to self-ascribe just consists in being able to ascribe properties to a particular object, oneself, thought of in first-personal terms. Noting the equivalence doesn't itself establish that our best theories of the mind could be properly expressed without the first-person. Nor does it show us where to look for an account of how creatures might become able to think first-personally, offer any explanation of the distinctive cognitive and epistemic properties of such thought, and so on. But those are surely interesting philosophical and psychological questions in their own right.

There are similar worries about Grant's argument against Yablo. What is it to 'access an ensemble of worlds'? For Grant's argument to go through, it had better be that we have accessed the metaphorical content in the relevant way when we have identified a uniquely specifying description of it. But that approach to grasp of content seems implausible in general. There is a good sense of access in which somebody with normal color vision has a way of accessing the colors of things that is unavailable to one who lacks it. But the latter subject can identify a uniquely specifying description of the relevant set of objects: namely, the set S of objects that the former subject's distinctively visual concept of blue correctly applies to. If cognitive access only required being able to find some way of uniquely specifying the set of possibilia that an expression was correlated with, then color vision would not be an explanatorily privileged way of accessing the colors of things. But there is a good sense in which the color-blind cannot access the colors of things, and so there is a good sense of cognitive access which requires more than unique specification of content.

If the above analogies are compelling, then settling the question of the indispensability of linguistic metaphor may require us to understand what is involved in thinking metaphorically. That second, broader project, however, may not be one that falls within the scope of philosophy of language as traditionally conceived.

References

Bergmann, M. 1982. "Metaphorical assertions." *Philosophical Review*, 91(2): 229–245.

Bezuidenhout, A. 2001. "Metaphor and what is said: a defense of a direct expression view of metaphor." *Midwest Studies in Philosophy*, 25(1): 156–186.

Camp, E. 2006a. "Metaphor and that certain 'Je ne sais quoi.'" *Philosophical Studies*, 129(1): 1–25.

Camp, E. 2006b. "Metaphor in the mind: the cognition of metaphor." *Philosophy Compass*, 1: 154–170. DOI:10.1111/j.1747-9991.2006.00013.x.

Carston, R. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Carston, R. and C. Wearing. 2015. "Hyperbolic language and its relation to metaphor and irony." *Journal of Pragmatics*, 79: 79–92.

Davidson, D. 1979. "What metaphors mean." In *On Metaphor*, edited by S. Sacks, pp. 29–45. Chicago: University of Chicago Press. Also reprinted in *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 1984, pp. 245–264.

Denham, A. 2000. *Metaphor and Moral Experience*. Oxford: Oxford University Press.

Glanzberg, M. 2007. "Metaphor and lexical semantics." *The Baltic International Yearbook of Cognition, Logic and Communication*, 3(1): 1–47.

Grant, J. 2010. "The dispensability of metaphor." *British Journal of Aesthetics*, 50(3): 255–272.

Leezenberg, M. 2001. *Contexts of Metaphor*. Amsterdam and New York: Elsevier.

Lepore, E., and M. Stone. 2010. "Against metaphorical meaning." *Topoi*, 29(2): 165–180.

McGonigal, A. 2008. "Davidson, metaphor and error theory." In *New Waves in Aesthetics*, edited by K. Stock and K. Thomson-Jones, pp. 58–83. London: Ashgate.

Peacocke, C. 2009. "The perception of music: sources of significance." *British Journal of Aesthetics*, 49(3): 257–275.

Recanati, F. 2004. *Literal Meaning*. Cambridge and New York: Cambridge University Press.

Reimer, M. 2001. "Davidson on metaphor." *Midwest Studies in Philosophy*, 25(1): 142–156.

Sperber, D., and D. Wilson. 1995. *Relevance: Communication and Cognition*, 2nd edn. Oxford: Blackwell.

Walton, K. L. 1993. "Metaphor and prop oriented make-believe." *European Journal of Philosophy*, 1(1): 39–57.

Yablo, S. 1998. "Does ontology rest on a mistake?" *Proceedings of the Aristotelian Society*, suppl. vol. 72: 229–226.

Further Reading

Hills, D. 1997. "Aptness and truth in verbal metaphor." *Philosophical Topics*, 25(1): 117–153.

Lewis, D. 1979. "Scorekeeping in a language game." *Journal of Philosophical Logic*, 8(1): 339–357.

Sperber, D., and D. Wilson. 2008. "A deflationary account of metaphors." In *The Cambridge Handbook of Metaphor and Thought*, edited by R. W. Gibbs, Jr, pp. 84–105. Cambridge and New York: Cambridge University Press.

Stern, J. 2000. *Metaphor in Context*. Cambridge, MA: MIT Press.

17

Conditionals

ANTHONY S. GILLIES

1 Introduction: Conditional Information

I want very much that you have the information that the beer is gone. In fact, I want you to take action that requires it. (Let us also stipulate that that action is unavailable (insert your favorite constraint here: policy, prudence, politeness) if there is plenty of beer.) So I say something that gives voice to the state of affairs, beer-wise:

(1) The beer is gone.

I hope that you understand me and take that information on board and then do the right thing. But things won't work out if I'm mistaken about the facts, for then what I am trying to pass off as information isn't that and my hopes will go unfulfilled. Nutshell: successful information exchange depends on the way things are.

That's clear (enough) if the information I aim to get to you is plain (enough) and the linguistic vehicle simple (enough). Conditional information is a useful kind of information and it is no surprise that natural language has canonical ways of expressing it. That is what *if*s – conditionals – are for. So take a case of *conditional* information exchange where the information is less plain and the linguistic vehicle less simple:

(2) a. If Jimbo is here, then he bought this round.
b. If Jimbo is here, then he might buy this round.
c. If Jimbo were here, then he would buy this round.

The information at stake here is information about what is or might or might not be the case *if* Jimbo is here, and what would or wouldn't be the case *if* he had been. Nutshell: successful (conditional) information exchange depends on the way things are but also on the ways things are in various alternative scenarios (the way things are if such-and-such).

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Not all conditional information is the same, and this is reflected in differences in conditionals. Take a so-called *Adams pair* (Adams, 1975):

(3) a. If Oswald didn't kill Kennedy, then someone else did.
b. If Oswald hadn't killed Kennedy, someone else would have.

The first conditional is an *indicative* conditional, saying what is the case if it turns out that Oswald *wasn't* involved. The second conditional is different. It has distinctive tense/aspect saying what *would have been* if, contrary to the facts, Oswald *hadn't* been involved. Such conditionals are *counterfactual* conditionals, but the name isn't a perfect fit.[1] The types of conditionals are genuinely different since (3a) and (3b) have the same antecedent and the same consequent, but one is true and the other false.

Two not-unrelated bundles of questions will occupy us here, regardless of the type of conditional we are considering. First bundle: What is that conditional information that is conventionally carried by various conditional constructions in natural language? Answering this will involve saying something about a conditional's dependence both on how things are and on how things might have been. It will also (but not invariably as we will see) involve saying when conditionals are true and when they are false. Second bundle: How do the various conditional constructions in natural language manage to carry that information? Answering this will involve saying something both about how a conditional's meaning arises from the parts of it and about how a conditional's meaning interacts with and contributes to the meaning of embedded and embedding environments. The bundles are in principle separable, but in practice often enough go hand in hand.[2] The first thing is all about saying what conditionals mean (what semantic values they have) and how things that mean those things are used in well-run conversations (what their pragmatic profiles are). The second thing is all about how *if* interacts with the rest of our language – how what specific conditionals mean is determined by the bits that make them up and about how conditional constructions contribute whatever-it-is they mean to embedding environments in which they occur as proper parts.

The aspiration is to at least see where some answers can be found.[3]

2 Preliminaries

Distinguish between (i) conditional sentences (indicative or counterfactual) of natural language and (ii) conditional connectives of some formal language that serves to represent the relevant conditional sentences. The aim is to associate an *if*-in-a-natural-language with an *if*-in-the-formal-language that then gets associated with its semantic value. We will assume that the role of the formal language can be adequately played by a simple propositional language with the usual sentential connectives ($\neg$, $\wedge$, $\vee$, $\supset$) plus a binary sentential connective (*if* $\cdot$)($\cdot$). Context will disambiguate which sort of conditional (*if* $\cdot$)($\cdot$) represents.[4] (When the time comes, we will also have use for modal operators of the usual sort like $\Box$ and $\Diamond$ and perhaps a few other things.) This two-step route to assigning meanings to conditionals isn't obligatory and is in principle dispensable (assuming that the mapping from natural language to the formal language (we won't fuss with this one) and then the mapping from the formal language to the universe of meanings (we will fuss with this one) are well behaved). But it does make for a clearer view of the landscape.

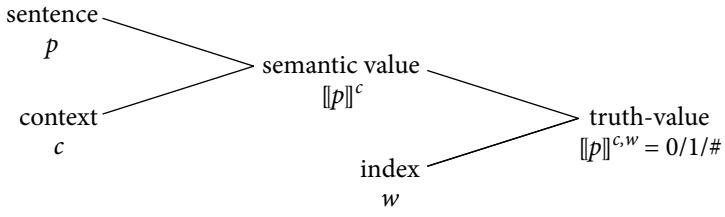

Figure 17.1 Variable but simple semantic values.

A (formal) language is only useful if it is interpreted. Whether or not truth-in-English has been achieved by an utterance of

(4) Jimbo has to be washing the dishes.

depends on the facts – how things actually are at the world where it is uttered. It also depends on the state of the conversation when it is uttered: for instance, whether what is being claimed is a claim about Jimbo's obligations or about what we know about his current activities. And it also depends on how things are in various other situations: for instance, whether at those relevant situations the prejacent *Jimbo is washing the dishes* is true. So truth-in-English is sensitive to both a context of utterance and an index of evaluation. One of the things we're after is a way of systematically saying when a sentence p is true in a context at a world.[5] So let's suppose that semantic values (whatever they are) determine truth-values at points of evaluation with respect to contexts. We will not need to say just what contexts are. It will be enough to carve out what role they play in the setup. Similarly, our indices will be worlds, but the most we need to know about them is that they are the kinds of things at which sentences are true or false.[6] The setup is summarized in Figure 17.1: a sentence p in a context c gets associated with a semantic value $[\![p]\!]^c$, which combines with an index w to deliver a truth-value (if the sentence has one) of the sentence at that world in that context, $[\![p]\!]\ c,w = 1$ or $[\![p]\!]\ c,w = 0$ or $[\![p]\!]\ c,w = \#$ (as the case might be). These aren't entirely innocent assumptions, but we can start here and suspend them when it suits us.

Theories of conditionals are constrained by the patterns of intuitive entailment they participate in. So what-intuitively-entails-what is important data and we want to explain its patterns as best we can. For conditionals this might tie what $(\mathit{if}\ \cdot)(\cdot)$ might mean very tightly to what we can and can't say about entailment. For instance, take the basic deduction theorem.

Deduction theorem $X, p \vDash q$ iff $X \vDash (\mathit{if}\ p)(q)$

Taking this as a constraint, what we say about $(\mathit{if}\ \cdot)(\cdot)$ impacts what we can say about $\vDash$ and vice versa. (This holds for various weakenings of the deduction theorem too, wherein the weakness lies in how the set of premises A and the premise p are combined.) The point is just that "entailment" is as much part of the theoretical machinery as is anything. So it's up for grabs whether what we say about *if*s bends to the will of entailment or whether what we say about entailment bends to the will of the *if*s. There is very little on stone tablets. That said, let's try to skirt issues about entailment when we can.

The simplest (and in a precise sense the weakest) conditional is the material conditional: $p \supset q$ is true iff either p is false or q is true. Accordingly, the simplest (and in a precise sense the weakest) theory of *if* takes it to simply be the material conditional.

Definition 1 (Horseshoe theory[7]). Conditionals are material conditionals:

$$[\![(\mathit{if}\ p)(q)]\!]\,c,w = 1 \text{ iff either } [\![p]\!]\,c,w = 0 \text{ or } [\![q]\!]\,c,w = 1.$$

This is, so far, neutral about whether the target conditionals are indicative or counterfactual. That matters and we will return to it.

In any case, this treats conditionals as truth-functional (and, by the way, context-invariant): the truth-value of a conditional at a world is entirely determined by the truth-value of its antecedent and consequent at that world. In fact, this is the only truth-functional option available for the conditional.

Fact 1. If $(\mathit{if}\ \cdot)(\cdot)$ is truth-functional then $(\mathit{if}\ \cdot)(\cdot) = \supset$.

Here's why. Suppose p = *The die came up six* and q = *the die came up even*. Our theory has to then render $(\mathit{if}\ p)(q)$ true no matter how the roll came out. So, in particular it's true if it came up six (antecedent true, consequent true), if it came up four (antecedent false, consequent true), and if it came up three (antecedent false, consequent false). The rub is that since by hypothesis $(\mathit{if}\ p)(q)$ only depends on the (actual) truth-values of p and q, any p and q with the same truth-values can be substituted in for them and the resulting conditional has to still be true. So: if a truth-functional conditional has a false antecedent or true consequent it is true. Now assume some conditional is false. It can only be because it has a true antecedent and false consequent. Given truth-functionality, this then holds for all conditionals. So: if a truth-functional conditional is false it has a true antecedent and false consequent.

As a theory of counterfactuals, this is an obvious non-starter. Conceptually, counterfactuals – the real deal ones with false antecedents – ask us to consider other, non-actual ways things might have been. But a truth-functional theory says that such considerings can't be relevant. The empirical coverage is also terrible. At least some counterfactuals with false antecedents are (contingently) true and thus worth arguing about. The material conditional rules that out.

(5) a. If Alex had come to the party, she would have arrived before 8.
b. If Alex had come to the party, she wouldn't have arrived before 8.

These conditionals cannot both be true, and speakers using them seemingly disagree. That is not what you'd expect if counterfactuals were horseshoes.

When it comes to indicative conditionals things are different: the horseshoe isn't widely adopted, but it is not without defenders. The main difficulty is that the material conditional is, from a logical point of view, weak. Taking indicatives to be horseshoes thus predicts that there are more entailments to conditionals than there seem to be. Among them: the paradoxes of material implication.

(6) a. Carl came alone.
??So: if Carl came with Lenny, neither came.
b. Billy got here first.
??So: if Alex got here before Billy, Billy got here first.

These don't strike as entailments even though the truth of either $\neg p$ or q at w secures the truth of $p \supset q$ at w.

The thing that has to be said is that while these are genuine entailments, there are pragmatic reasons – derived from how conditionals and surrounding sentences are reasonably and appropriately deployed in conversation – why they strike us as weird. For instance: the

conditional conclusions in (6) are weird because in each case a speaker in the position to assert the unconditional premise has no use for (and hence would mislead by using) the logically weaker conditional conclusion. So there is a clash between an implicature of the conclusion and the initial premise.[8]

Explaining away unwanted entailments by appeal to implicatures is tricky since this strategy has nothing to say about conditionals that occur unasserted in embedded environments. But conditionals occur in such environments and when they do the horseshoe theory makes some unhappy predictions. Negated indicatives are a case in point. The issue is that since material conditionals are so weak, their negations are correspondingly strong.

(7) It's not so that if the gardener didn't do it then the butler did.

I can be signed up for (7) without being signed up for the truth of *the gardener didn't do it*. After all, it might have been the driver. Maybe we need a pragmatic defense of our pragmatic defense. Perhaps what we have in (7) is a denial of a conditional rather than an assertion of a negated conditional, and so the negation is not a negation. I doubt it since the issue can be pushed where the negated conditional itself occurs embedded and thus unasserted and thus not open to this defense. For instance: the argument in (8) is a disaster but predicted to be an entailment by the horseshoe.

(8) If there is no god, then it is not the case that if I pray, my prayers will be answered.
I don't pray.
??So: there is a god.

Is there another pragmatic defense to save these pragmatic defenses of the earlier pragmatic defense? It's possible, but this rescue is quickly becoming a wheels-within-wheels situation.

3 Strict Conditionals

The material conditional is extreme in its myopia: only how things are at w matter to the truth of a material conditional at w. At the other end of the spectrum lie *strict conditionals*: these survey *all* possibilities. That is a quantificational claim and so, as with a lot of quantificational claims, this one may be restricted.[9] We will care about the *if*-relevant worlds in a context c. A strict conditional is a (restricted) universal quantifier, saying that all the *if*-relevant possibilities at which the antecedent is true are possibilities at which the consequent true. When the only job of contexts is to provide such worlds – as it is now – let's simply identify a context c with the selection function that delivers the *if* -relevant for each world w.

Definition 2 (Strict conditionals). Let $c(w)$ be the set of *if*-relevant worlds at w (in c).

$$[\![(\textit{if}\ p)(q)]\!]\,c,w = 1 \text{ iff if } v \in c(w) \text{ and } [\![\,p]\!]\,c,v = 1 \text{ then } [\![q]\!]\,c,v = 1.$$

This is equivalent to saying that $(\textit{if}\ p)(q)$ is true at w with respect to c iff the material conditional $p \supset q$ is true at every world in $c(w)$.

What does it take to be an *if*-relevant world? We haven't said (that's by design). Thus, depending on what we say about what it takes for a world to be an *if*-relevant world at w (that is, depending on what we say about the function c given an argument w), we get a different strict conditional.

Here are some possibilities (not exhaustive):

- For any w, only w ever matters (for every w: $c(w) = \{w\}$). The resulting strict conditional is the material conditional.
- For any w, all worlds (unrestricted!) always matter (for every w: $c(w) = W$). The resulting strict conditional is strict implication (true iff the antecedent entails the consequent).
- For any w, all worlds compatible with what is known in w matter (for every w: $c(w) = \{v$: if X is known at w then $v \in X\}$). The resulting strict conditional is a sort of epistemic strict conditional.
- For any w, all worlds similar to w to fixed degree d matter (for every w: $c(w) = \{v : v$ is similar to w to at least degree $d\}$). The resulting strict conditional is a sort of similarity-based strict conditional.

Notice that for each such $c(w)$ there is a corresponding (restricted) necessity operator $\Box_{c(w)}$ and a strict conditional $(if\ p)(q)$ with respect to $c(w)$ amounts to claiming that the corresponding material conditional $p \supset q$ is $\Box_{c(w)}$-necessary. The strictness of two strict conditionals can be compared: if the set of *if*-relevant worlds (at a world) for one strict conditional is included in the set of *if*-relevant worlds for a second strict conditional, then since the second quantifies over more possibilities it is a stricter strict conditional.

Fact 2. Suppose $(if_1\ p)(q)$ is a strict conditional wrt $c_1(w)$ and $(if_2\ p)(q)$ is a strict conditional wrt $c_2(w)$. If $c_1(w) \subseteq c_2(w)$ then $(if_2\ p)(q)$ implies $(if_1\ p)(q)$.

This way of thinking about conditionals treats them as restricted necessity operators, saying that their consequents are true throughout a given set of antecedent worlds. Since we can probe necessity operators by examining the properties of the functions which determine the sets of worlds they quantify over, the same is true of restricted necessity and so of strict conditionals. For instance: suppose that every world is relevant to an *if* at that world. That means that c is *reflexive*: for every w it is the case that $w \in c(w)$.

Fact 3. c is reflexive iff $(if\ p)(q)$ implies $p \supset q$.

Two comments. First: this amounts to saying that *if* is on the spectrum of logical strength, occupying a place *no weaker* than the material conditional. (Since $c(w)$ is always a set of worlds we have built-in that strict implication is at least as strong as *if*.) Since the material conditional together with its antecedent entails its consequent (on any sensible notion of "entails") it follows that being stricter than the material conditional means going in for some version of *modus ponens*. We will come back to this. Second: this follows straightaway from the fact that $(if\ p)(q)$ with respect to $c(w)$ is equivalent to $\Box_{c(w)}(p \supset q)$ and the fact that in modal logic reflexive accessibility relations (or functions or spheres or whatever) correspond to the validity of instances of the T axiom $\Box p \supset p$.

Another sort of property is the package of import/export equivalences.

Import/export $(if\ p \wedge q)(r) \dashv\models (if\ p)((if\ q)(r))$

Some potential import/export pairs for indicatives:

(9) a. If Carl is away, then if Lenny is away, then Sector 7G is empty.
b. If Carl is away and Lenny is away, then Sector 7G is empty.

And for counterfactuals:

(10) a. If Alex had come to the party, then if Billy had come to the party it would have been great fun.
b. If Alex had come to the party and Billy had come to the party, it would have been great fun.

Once we are thinking of conditionals as restricted necessity operators, import/export reads a lot like some sort of constraint on introspection. That's close to right.[10]

Fact 4. $(\mathit{if}\,\cdot)(\cdot)$ supports the right-to-left direction of import/export iff c is *shift reflexive*. That is: iff $v \in c(w)$ implies $v \in c(v)$. It supports the left-to-right direction iff c is *shift coreflexive*. That is: iff $v \in c(w)$ and $v \in c(u)$ implies $u = v$.[11]

Together (shift) reflexivity/coreflexivity – that is, both halves of import/export – impose a very strong requirement: that whenever a world is *if*-relevant at w that world is related to itself and to no other. (This property is what Kaufmann and Kaufmann call "shift identity.") That seems to be a weird requirement.

Matters get immediately worse. Combining reflexivity of the assignment of *if*-relevant possibilities with shift reflexivity/coreflexivity means that that assignment is an island function: every world is an *if*-relevant island to a conditional at that world (for every w: $c(w) = \{w\}$).

Fact 5. $(\mathit{if}\,\cdot)(\cdot)$ implies the corresponding material conditional and supports import/export iff for every w: $c(w) = \{w\}$.

As we saw: a strict conditional over such island sets of *if*-relevant possibilities is just the material conditional. So any conditional comparable to and no weaker than material implication that supports import/export must be the material conditional. Such "collapse" arguments serve as clear maps to the landscape of conditionals (especially indicatives). We will return to this theme.

There are other properties that strict conditionals (of whatever strictness) have, irrespective of the properties of the sets of *if*-relevant possibilities. I will mention three.

Antecedent strengthening $(\mathit{if}\,p)(q) \vDash (\mathit{if}\,p \wedge r)(q)$ for any r
Transitivity $(\mathit{if}\,p)(q), (\mathit{if}\,q)(r) \vDash (\mathit{if}\,p)(r)$
Contraposition $(\mathit{if}\,p)(q) \vDash (\mathit{if}\,\neg q)(\neg p)$

Fact 6. Any strict conditional supports antecedent strengthening, transitivity, and contraposition.

The reason for all three properties is simple: strict conditionals are universal quantifiers, saying all the *if*-relevant possibilities are of a certain sort. And the thing about *all*: all-claims are downward monotone, support transitive inferences, and contrapose.[12] Some quick Venn diagrams will convince the unmoved.

These properties push away from strict conditional analyses of (especially) counterfactuals, and (as we'll see) predicting that these patterns are patterns of non-entailment is a highlight of the classic *variably strict* theory developed initially by Stalnaker (1968; 1984) and Lewis (1973). Some concrete counter-examples:

(11) a. If Sophie had gone to the parade, she would have seen Pedro dance.
??So: If Sophie had gone to the parade and been stuck behind someone tall, she would have seen Pedro dance.
b. If Hoover had been a communist, he would have been a traitor.
If Hoover had been born a Russian, he would have been a communist.
??So: If Hoover had been born a Russian, he would have been a traitor.
c. If it had rained, it wouldn't have poured.
??So: If it had poured, it wouldn't have rained.

In fact the indicative counterparts here also make trouble for strict conditional accounts of indicatives. Set this aside for now.

4 Variably Strict Conditionals

We have come this far without saying whether the *if*s at stake are indicatives or counterfactuals. Let's change that by focusing on counterfactuals (though, not quite exclusively). The classic account of counterfactuals, and by far the account that remains dominant, treats them as *variably strict conditionals* (Stalnaker, 1968; Lewis, 1973).

Look again at (11a). This is problematic for strict conditionals since it seems eminently possible that (i) had Sophie gone she would have seen Pedro dance, and (translating to our formal language) so (ii) the conditional $(\mathit{if}\ p)(q)$ is true. Though, admittedly, (iii) not if she had been stuck behind someone tall. So, (iv) the counterfactual $(\mathit{if}\ p \wedge r)(q)$ is false and, it seems, (v) the contrary conditional $(\mathit{if}\ p \wedge r)(\neg q)$ is true. Strict conditionals don't seem to have the needed flexibility to allow both (ii) and (v). Put another way: sequences of counterfactuals like $(\mathit{if}\ p)(q)\,;(\mathit{if}\ p \wedge r)(\neg q)$ seem like they can be consistent – both conditionals can be non-vacuously true at the same time. Since that's not true of strict conditionals, Lewis (1973) argued on this basis that counterfactuals can't be any strict conditional no matter the level of strictness.

Some examples:

(12) a. If the USA threw its weapons into the sea tomorrow, there would be war;
b. But if the USA and the other nuclear powers threw all their weapons into the sea tomorrow, there would be peace.

(13) a. If you had been standing a foot to the left, you would have been killed;
b. But if you had (also) been wearing your hard hat, you would have been alright.

(14) a. If Sophie had gone to the parade, she would have seen Pedro dance;
b. But if Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro dance.

With a little ingenuity, it seems each sequence can be extended for as long as you like, with each successive conditional having an ever-stronger antecedent and the consequent flipping from negated to not. Such sequences are called *Sobel sequences*.[13] Analyzing any Sobel sequence as a sequence of strict conditionals, no matter the strictness, won't work and the reason is just that strict conditionals support antecedent strengthening.

Lewis's conclusion is that there is no level of strictness such that in a given situation counterfactuals are strict conditionals of that strictness. Instead, they are variably strict

conditionals: they are more or less strict depending on the strengths of their antecedents. Intuitively (with a not insubstantial additional assumption that we'll return to) a counterfactual like (14a) says that all of the possibilities most similar to ours but in which Sophie went to the parade are possibilities wherein she witnessed Pedro's dancing. The "thinned" (14b) on the other hand says something about all of the possibilities most similar to ours but in which Sophie went to the parade *and* was stuck behind someone tall (namely that in those worlds she missed out on the dancing). These are compatible so long as the (modally) nearest go-to-the-parade possibilities do not include the nearest go-to-the-parade-and-get-stuck-behind-someone-tall possibilities.

To make this less impressionistic: rather than simply associating with each world w a set $c(w)$ of *if*-relevant possibilities – possibilities relevant regardless of the *if* – we want to look at the nearest or most similar possibilities to w that make a particular antecedent true. So we need some (context-dependent) measure of relative similarity or nearness or closeness of worlds to w. We'll write it this way: $\leq_w$, with the idea that if $v \leq_w u$ then v is at least as close to w as u is.[14] Officially we are thinking that the context c determines such an ordering for each world, but we won't say how that happens and will generally suppress mention of c. Even before saying anything about this relation we know how the story will appeal to it.

Definition 3 (Variably strict conditionals). Conditionals are variably strict conditionals:

$$[\![(\mathit{if}\ p)(q)]\!]\,c,w = 1 \text{ iff } \text{if } v \text{ is a} \leq_w\text{-minimal } p\text{-world then } [\![q]\!]\,c,v = 1$$

where a world v is a $\leq_w$-minimal p-world iff there is no p-world u such that $u <_w v$.

Given this setup, there are some straightforward things to be said about the nearness ordering. For one thing, it has to actually *be* an ordering. For another thing, it should respect the fact that each world is an awful lot like itself and so pretty close to itself.

Definition 4 (Relative similarity). For any worlds w, v let $\leq_w$ be such that:

(i) (Ordering Assumption) $\leq_w$ is a connected and transitive relation over W.
(ii) (Centering Assumption) w is minimal in $\leq_w$: if $v \leq_w w$ then $w = v$.

Together Definition 3 and Definition 4 form the bare bones of the story.[15]

There are other notable assumptions we might add to it by imposing further constraints on the ordering (for any proposition X and world w):

(iii) (Limit Assumption) X has at least one $\leq_w$-minimal world.
(iv) (Uniqueness Assumption) X has no more than one $\leq_w$-minimal world.

Stalnaker's (1968) setup makes both of these additional assumptions, while Lewis's (1973) makes neither. Some of the properties are directly reflected in logical principles. For instance:

Conditional excluded middle $\models (\mathit{if}\ p)(q) \vee (\mathit{if}\ p)(\neg q)$

Fact 7. If $\leq_w$ satisfies both the Limit Assumption and the Uniqueness Assumption then $(\mathit{if}\ \cdot)(\cdot)$ supports conditional excluded middle.

The additional assumptions on the orderings are separable. I won't be assuming uniqueness (for the most part) but will make the limit assumption – indeed, above in Definition 3 I already made the Limit Assumption.[16]

It is easy to see why this approach is attractive. Take our Sobel sequences (12)–(14). If the conditionals are variably strict then no wonder they are consistent: in each case the first conditional takes us to the nearest possibilities in which a simple antecedent is true, but the second one takes us to a set of possibilities in which *in addition* some other thing is true. Since these need not be the same (indeed, aren't), the two conditionals end up talking about different possibilities.

The same goes for the problematic entailments that strict conditionals support.

Fact 8. Variably strict conditionals do not support antecedent strengthening, transitivity, or contraposition.

Variably strict conditionals don't validate any of these patterns and that is in their favor as a theory of counterfactuals in natural language.

Given our assumptions, we could easily retell the story in terms of sets of (antecedent specific) *if*-relevant worlds rather than the (family of) nearness or similarity orderings by, for a given world w and antecedent p, collecting up the set of worlds no further from w than the nearest p-world.

Definition 5 $c(w) = \lambda p\,.\{v$: if u is a minimal p-world in $\leq_w$ then $u \not<_w v\}$

One thing to notice is that centering (actually, weak centering) translates to the constraint that for any w and (consistent) p it is the case that $w \in c(w)(p)$. That means that counterfactuals would support *modus ponens*, and *that* seems strange since counterfactuals carry some sort of signal that their antecedents are false and those aren't the kinds of things that *modus ponens* comes up for.[17] Still, I think this is as it should be.

(15) a. Alex: If my keys had been in the drawer, I would have seen them.
b. Billy: No, look: they were in the drawer under the mail and you missed them!

Billy's denial here makes sense because the truth of the counter-example ($p \wedge \neg q$) at a world is always sufficient for the falsity of the counterfactual (*if* p)(q) at that world. And that is what (weak) centering imposes.

So we have reflexivity. Since variably strict conditionals are stronger than material conditionals (i.e., the sets of relevant *if*-worlds aren't always trivial islands) we know that variably strict conditionals can't validate import/export.

Fact 9. Variably strict conditionals do not support import/export.

Import/export would require that the closest q-worlds to the closest p-worlds are the same as the closest ($p \wedge q$)-worlds. Counter-examples (of the formal stripe) are easy to construct. For instance: suppose we have just the worlds w, x, y, z where p is true at x and y while q is true at y and z. Now, let $\leq_w$ be the ordering $w \leq_w x \leq_w y \leq_w z$. Thus the ($\leq_w$-)nearest ($p \wedge q$)-world to w is y. Finally, let $\leq_x$ be such that $z <_x y$. Thus the ($\leq_x$-)nearest q-world to the ($\leq_w$-)nearest p-world to w isn't y.

As we saw with conditional excluded middle (and the Limit Assumption) properties of the nearness relations (metaphor!) make for differences in the properties that an analysis of

conditionals based on them has. But disagreements about the properties of the nearness relation can seem to be bigger than they in fact are. Assume for simplicity that the space of worlds is finite. Then there are two natural decision points. First decision point: does the ordering permit ties? That is, is the case that two distinct worlds u and v can be tied in closeness or similarity to a world w? Adopting the Uniqueness Assumption prohibits ties. Second decision point: must all worlds always be comparable? That is, (for any w, u, v) is it the case that either $u \leq_w v$ or $v \leq_w u$? In principle these are separate questions, but in practice the dividing lines are whether ties and incomparabilities are both prohibited (Stalnaker), whether ties are permitted but incomparabiltities are not (Lewis), or whether both ties and incompatibilities are permitted (Pollock).

But, as Lewis (1981) showed, there is less disagreement here than appears. Suppose our ordering permits neither ties nor incomparabilities. That's true for any *fixed* and *determinate* ordering and, for all that we've said so far, it's possible that ordinary contexts don't always determine unique orderings. In such contexts there are, let's say, multiple admissible orderings. Then supervaluating over the admissible orderings gives determinate truth-conditions for simple conditionals just like we would get from a theory that permits ties but not incomparabilities.[18] And similarly for the second choice point. A theory that permits ties but not incomparabilities but allows that there may be a range of admissible orderings in any given context is, in this same sense, equivalent to a theory like Pollock's (1976) which permits both ties and incomparabilities.

There is one more set of equivalences between theories to round out this part of the discussion. One way to make sense out of the idea of conditional information is embodied by (some version of) the Ramsey test.

> **Ramsey test** (*if p*)(*q*) is true/accurate/acceptable in a situation iff *q* is true or accurate or acceptable in that situation-plus-the-information-that-*p*.

Depending on what we say about "true/accurate/acceptable," "situations," and "situation-plus-the-information-that-*p*" we get different theories.[19]

There are ways of filling in those details for counterfactuals. Most prominently among them are versions of *premise semantics*.[20] The idea is that (relative to a context, just as with the orderings) we associate with each world a set of propositions true at that world, a premise set P_w at w, and use that as the factual background to assessing a counterfactual's truth at w. Just like with similarity orderings, this assignment depends on context, though we won't say just how. A conditional (*if p*)(*q*) is true at a world iff every way of consistently adding the information that p to those premises gives us a derived or subordinate set of premises that entails q.

Let's say things a little more precisely (still sticking to the assumption that W is finite) so we can say how this setup relates to the variably strict picture.

> **Definition 6** (Premise sets). For every world w there is a set P_w of propositions such that:
>
> (i) (Exhaustive Assumption) Every world belongs to some proposition in P_w ($\cup P_w = W$).
> (ii) (Centering Assumption) Every member of P_w is true at w and w only ($\cap P_w = \{w\}$).
>
> For any proposition X let $\mathbf{P}_w^X$ be the set of maximal X-consistent subsets of P_w.[21]

The first assumption plays the role of the Ordering Assumption for variably strict conditionals (our orderings left out no worlds) and the second assumption plays the role of centering (naturally enough).

Counterfactuals ask us what follows from the minimal ways of making room for the antecedent. So we look to the maximal p-consistent subsets of our premise set and see what follows from them.

Definition 7 (Conditionals and premise sets). Conditionals are minimally revised premise set entailments:

$$[\![(\mathit{if}\ p)(q)]\!]\ c,w = 1 \text{ iff if } P^{\star}_{w} \in \mathbf{P}\ w\ [\![p]\!] \text{ then } P^{\star}_{w} \text{ entails } [\![q]\!].$$

Lewis (1981) showed that such a theory is equivalent to a theory based on orderings that permits ties and incomparabilities between worlds. There are ways of turning premise sets into an ordering and the result of doing that gives us an ordering that says a counterfactual is true iff our original premise set did. (Going the other way, every ordering can be derived in this way from some premise set, with the same result.) It is the deriving of orderings from premise sets that is key:

Fact 10. A premise set P_w induces an ordering $\leq^{P}_{w}$ where $u \leq^{P}_{w} v$ iff every proposition in P_w true at v is true at u (that is: iff $\{X \in P\ w\colon u \in X\} \subseteq \{X \in P\ w\colon v \in X\}$). A conditional $(\mathit{if}\ p)(q)$ is true with resect to P_w iff it is true with respect to $\leq^{P}_{w}$.

So premise semantics and ordering semantics can both be sensibly seen as different but broadly equivalent implementations of the variably strict theory.[22]

5 Counterfactual Dynamics

Taking counterfactuals to be variably strict conditionals is the industry standard. But it is open to challenge.[23] A lot of weight has been put on the basic logic that variably strict conditionals deliver. But things are more subtle, and way more interesting, than they at first appear.

Look again at our examples of antecedent strengthening gone wrong (12)–(14). These bear important weight in the argument against taking counterfactuals to be any strict conditional. If they were any strict conditional, each pair would be incompatible, but they are as normal as counterfactual discourse ever gets. But as normal as they are, they are dramatically not that – indeed they seem contradictory – when issued in reverse order as *reverse Sobel sequences*.[24] Look at the difference with Lewis's example. Here is the (a)→(b) direction we saw before:

(12) a. If the USA threw its weapons into the sea tomorrow, there would be war;
b. But if the USA and the other nuclear powers threw all their weapons into the sea tomorrow, there would be peace.

And the reversed (b)→(a) direction:

(16) ??If the USA and the other nuclear powers threw all their weapons into the sea tomorrow, there would be peace; but if the USA threw its weapons into the sea tomorrow, there would be war.

Another of our (a)→(b) examples and its reversed (b)→(a) counterpart:

(14) a. If Sophie had gone to the parade, she would have seen Pedro dance;
b. But if Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro dance.

(17) ??If Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro dance; but if Sophie had gone to the parade, she would have seen Pedro dance.

There's no doubting the asymmetry here: unfolding a Sobel sequence (and so a counter-example to antecedent strengthening) from (a)→(b) is one thing, reversing the order from (b)→(a) is another thing entirely. But, from the variably strict point of view, once a counter-example to antecedent strengthening always a counter-example. Fiddling with the order in which we find the nearest worlds in which the USA throws its weapons into the sea tomorrow (maybe not so close) versus finding the nearest worlds in which the USA *and* all the other nuclear powers throw their weapons into the sea tomorrow (even further away) shouldn't change the fact that the sets we get are different in those different cases. This is puzzling.

The argument against treating counterfactuals as strict conditionals that is based on Sobel sequences isn't airtight. Sobel sequences show that counterfactuals aren't strict conditionals so long as the level of strictness never changes as counterfactual discourse unfolds. So there's an unargued-for assumption. Reverse Sobel sequences give us a reason to explore life without it.[25]

I will sketch a version of the strict conditional theory initially developed by von Fintel (2001) and then (not quite conservatively) extended in Gillies 2007. There are three ingredients to the story.[26]

Context keeps track of what worlds are counterfactually relevant.[27] Begin with a system of spheres around a world w: a nested set of sets of worlds $c(w)$ such that $\{w\} \in c(w)$ and $\cup c(w) = W$. Assume that context keeps track of such a set and that it determines the counterfactually relevant worlds.

Definition 8 The default *hyperdomain* at w in c is $\pi_c^0 = c(w)$. Given a hyperdomain π_c the *domain* of counterfactually relevant worlds (a.k.a. the *modal horizon*) s_{π_c} is the $\subseteq$-minimal domain in π_c.

So the default modal horizon at w is $s_{\pi^0} = \{w\}$. (When the context is clear (or not relevant since the counterfactual domains are the only things we are keeping track of here) we will omit reference to c.)

Conditionals presuppose that their antecedents are compatible with the relevant domain that the conditionals talk about. So counterfactuals presuppose that their antecedents are compatible with the counterfactual domain.

Entertainability presupposition A counterfactual (*if p*)(*q*) in context *c* presupposes that s_{π_c} contains some *p*-worlds.

Sometimes an assertion can be successful even if it has unmet presuppositions. That is because sometimes (*ceteris paribus* and within certain limits) we accommodate the missing presupposition and evaluate the assertion not with respect to the context as it was when the utterance was made but in the context changed to accommodate the missing presupposition.[28] And so it is with counterfactuals and their entertainability presuppositions. Asserting a counterfactual may change the conversational score when the presupposition is accommodated.

Definition 9 (Context change via accommodation). Let *c* be any context and *p* any counterfactual antecedent. Then $c|p|$ is the posterior context that results from accommodating the entertainability of *p* into *c* where

$$\pi_{c|p|} = \{s \in \pi\, c\colon\ s \cap [\![p]\!] \neq \varnothing\}$$

Notice that if s_{π_c} – the $\subseteq$-minimal domain in π_c – is compatible with *p* then $\pi_{c|p|} = \pi$. Otherwise $\pi_{c|p|}$ is the result of eliminating inner sets of worlds from π_c until we get one that has some *p*-worlds in it.

Putting the pieces together:

Definition 10 (Dynamic strict counterfactuals v.1.0). Counterfactuals are dynamic strict conditionals that carry entertainability presuppositions:

$$[\![(\mathit{if}\, p)\,(q)]\!]\ c,w{=}1 \text{ iff } (s_{\pi_{c|p|}} \cap [\![p]\!]) \subseteq [\![q]\!]$$

This is equivalent to saying that the counterfactual is true iff the corresponding material conditional is $\Box s_{\pi_{c|p|}}$-necessary.[29] So a counterfactual (*if p*)(*q*) in a context *c* presupposes that *p* might have been the case and then asserts the strict conditional $\Box(p \supset q)$ with respect to the post-accommodation domain provided by *c*-changed-just-a-bit.

Accommodation works by admitting as relevant the worlds as close as the nearest antecedent world.[30] Since we begin by default with the counterfactual domain containing just the world of evaluation, this means that the dynamic theory captures the variably strict theory as a special case when counterfactual discourse stretches only one conditional long.

Fact 11. Suppose context *c* delivers a system of spheres S_w. Then a counterfactual (*if p*)(*q*) is true at *w* in *c* according to the variably strict semantics iff it is true at *w* in *c* according to the dynamic strict conditional theory.

Things are different for sequences of counterfactuals. The dynamic theory treats counterfactuals as context dependent and context affecting. Context dependent because the modal horizon is a function of context. Context affecting because what worlds are counterfactually relevant can be changed by a counterfactual's antecedent. It is this interplay that explains the asymmetry between (a)→(b) Sobel sequences like (12) and (14) and their (b)→(a) reverse Sobel sequences like (16) and (17).

Here's how. Suppose *p* is false at *w* but that the nearest *p*-worlds to *w* are *q*-worlds. For instance: suppose that Sophie did not go to the parade and Pedro was conspicuous and

reliably a dancer at the parade. The default domain $\{w\}$ doesn't satisfy the counterfactual's presupposition. So we accommodate. It's possible for all the p-worlds in $s_{c|p|}$ to be q-worlds and so for $(if\ p)(q)$ to be true.

Now we come to the second counterfactual in an (a)→(b) Sobel sequence. But the context is not the default initial context. It is $c|p|$. Still, it is possible that there are no r-worlds in $s_{\pi_{c|p|}}$. (Sophie is not overly tall and doesn't generally travel with stilts.) The presupposition of $(if\ p \wedge r)(\neg q)$ isn't met so we accommodate. It's possible for all the $(p \wedge r)$-worlds in $s_{\pi_{c|p\wedge r|}}$ to be $\neg q$-worlds: the nearest worlds where Sophie gets stuck behind someone tall can all be worlds where she doesn't see Pedro dance.[31] So the counterfactual $(if\ p \wedge r)(\neg q)$ can be true, too.

The (b)→(a) direction is different. If we *first* accommodate the antecedent of the (b) counterfactual $(if\ p \wedge r)(\neg q)$ we go straightaway to context $c|p \wedge r|$. Assume that indeed all the $(p \wedge r)$-worlds in $s_{\pi_{c|p\wedge r|}}$ are $\neg q$-worlds so that $(if\ p \wedge r)(\neg q)$ is true. There is now no room for the simple (a) conditional to be true in the context $c|p \wedge r|$. That is because its presupposition – that p is possible relative to $c|(p \wedge r)|$ – is satisfied. Hence: no accommodation. But we just said that all the $(p \wedge r)$-worlds in $s_{\pi_{c|(p\wedge r)|}}$ are $\neg q$-worlds. And so it's false that all the p-worlds in $s_{\pi_{c|(p\wedge r)|}}$ are q-worlds. So the simple (a) counterfactual can't be true.

Members of Sobel sequences end up quantifying over different domains; members of reverse Sobel sequences end up quantifying over the same domain. No wonder there is an asymmetry. Figure 17.2 shows how this works graphically. For those who want to see this in slightly more formal terms, we need one more definition.

Definition 11 Let $C_0; \ldots; C_n$ be a sequence of counterfactuals and a_i be the antecedent of the ith counterfactual. Such a sequence is *satisfiable* iff there is a world w and a system of spheres $c(w)$ such that: (i) $[\![C\,0]\!]\ c\,0,w = \ldots = [\![C\,n]\!]\ c\,n,w = 1$, where (ii) $\pi^0_{c_0} = c(w)$ and (iii) $c_{i+1} = c_i|a_i|$ (for each $i > 0$).

This just captures what we have been assuming in working through the Sobel sequences: that interpreting counterfactuals changes the context and those changes can matter when it comes to seeing when a sequence is consistent or not.

Fact 12. Sobel sequences $(if\ p)(q)$; $(if\ p \wedge r)(\neg q)$ are satisfiable according to the dynamic strict conditional theory. Reverse Sobel sequences $(if\ p \wedge r)(\neg q)$; $(if\ p)(q)$ are not satisfiable.

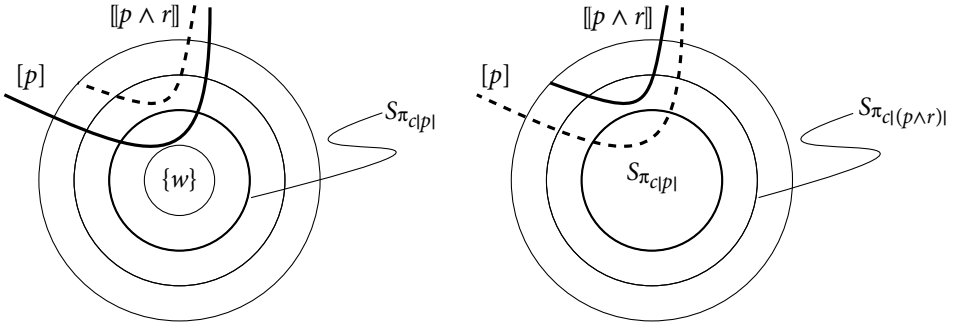

Figure 17.2 Accommodating p as a possibility (left) and $(p \wedge r)$ as a possibility (right).

The last word here has yet to be had. Counterfactual domains seem to get ever-larger. So one question: How do counterfactual domains get *reset*? Indeed there are examples (discussed in von Fintel, 2001, §8, and Gillies, 2007, §9) in which somehow or other the accommodation that a counterfactual's antecedent seems to trigger is promptly undone.

(18) If Sophie had gone to the parade and been stuck behind someone tall, she wouldn't have seen Pedro dance. Still, if Sophie had gone to the parade, she wouldn't have been stuck behind someone tall.

Some (for instance Moss, 2012) think the phenomenon tells decisively against the dynamic strict conditional theory. But there is room to move and the sort of thing that needs saying is clear: resetting happens, but it is not as smooth as expanding the counterfactual domain. There is evidence in favor of this, though the mechanism is poorly (i.e., not at all) understood. Some evidence that von Fintel discusses: modal and (certain) factual information makes the resetting smoother:

(19) If the USA threw its weapons into the sea tomorrow, there would be war. Well, if all the nuclear powers threw their weapons into the sea tomorrow, there would be peace. *But of course, that would never happen*. So, as things stand, if the USA threw its weapons into the sea tomorrow, there would be war.

(20) A: If John had been at the party, it would have been much more fun.
B: Well, if John had been at the party and had gotten into a fight with Perry, that wouldn't have been any fun at all.
A: *Yes, but Perry wasn't there*. So, if John had been at the party, he wouldn't have gotten into a fight with Perry.

Part of what makes this situation puzzling is that retreating to the variably strict theory does not seem like the right path. The reason has to do with *negative polarity items* (NPIs) like *any* and *ever* and when they are licensed. Some examples:

(21) a. Every diner who has *ever* eaten there has liked it.
b. Every diner who ordered *any* dessert loved it.
c. *Some diner who has *ever* eaten there has liked it.
d. *Some diner who ate there has *any* reason to go back.

Quantificational claims like these combine a quantificational determiner $\mathcal{Q}$ with a *restrictor* A and a *nuclear scope* B, saying that $\mathcal{Q}$-many of the As are B's. Some determiners (universal ones like *every*) are *downward entailing* in their first argument: *every* $(A)(B) \vDash$ *every* $(A')(B)$ if $A' \subseteq A$. And some (like *some*) are not: *some* $(A)(B) \nvDash$ *some* $(A')(B)$ for *some* $A' \subseteq A$. (In fact *some* isn't downward entailing in either argument.) Instead *some* is upward entailing in its first argument. Ladusaw (1980) was the first to argue that NPIs are licensed only when they are in a downward entailing environment. That elegantly predicts the distribution in (21). It is in fact a very robust generalization.

Counterfactual antecedents can also contain NPIs. This (as argued in von Fintel, 2001, and in von Fintel and Gillies, 2015) is trouble-making for any retreat to a variably strict theory.

(22) a. If you had eaten *any* fruit, you would have had an apple.
b. If you had *ever* been to Idaho, you would have liked it.

The trouble is that the antecedent of a variably strict conditional is not a downward monotone environment. Given our setup we can say something very stark.

Fact 13. A conditional $(\mathit{if}\ \cdot)(\cdot)$ is downward entailing in its antecedent iff it supports antecedent strengthening.

Antecedent strengthening is of course just what the variably strict conditional is designed to do without. Whence it leaves unexplained why counterfactual antecedents license NPIs. The variably strict conditional theory gets this wrong for exactly the reason the strict conditional theory gets it right.

6 Indicative Conditionals and Collapse

We know that logical space is limited for a conditional that is bounded by strict implication (from Fact 5) and material implication (from below) and supports import/export. Material conditionals seem to be the only option. So if you are not satisfied with the horseshoe as a theory of ordinary indicative conditionals, you have choices to make. They are not easy choices.[32]

Another collapse argument:

(23) Either the butler did it or the gardener did.
So: if it wasn't the butler, it was the gardener.

Patterns like this – the *direct argument* or more descriptively *or*-to-*if* – seem like entailments. But that makes for immediate trouble.

Fact 14. If $p \vee q \vDash (\mathit{if}\ \neg p)(q)$ then $p \supset q \vDash (\mathit{if}\ p)(q)$.

The proof is simple: suppose $p \supset q$ is true (at w in c). Then so is $\neg p \vee q$ and hence by *or*-to-*if* so is the indicative $(\mathit{if}\ p)(q)$. If indicatives are bound from below by material conditionals, then we have the stronger collapse conclusion: $(\mathit{if}\ p)(q) \dashv\vDash p \supset q$.

These seem like puzzles. Perhaps even puzzles-in-the-pejorative-sense (whatever that is). But the issue collapse arguments like these bring out is bigger than that. Indicative conditionals seem to be stronger than their corresponding material conditionals. They seem to express something more. But what this more is and how it gets expressed is not obvious. Collapse arguments like these are good ways of mapping logical space for how to think about this problem. But they also reveal that the going won't be easy (the assumptions they require are modest and super plausible).

So, one option is to deny the lower bound. If that is right then the truth of a counterexample $(p \wedge \neg q)$ isn't always sufficient for the falsity of the indicative $(\mathit{if}\ p)(q)$. And so, if indicatives do not entail their corresponding material conditionals, it is possible for $(\mathit{if}\ p)(q)$ and p to both be true, but for q to be false. It's not enough to just deny an assumption: what we need is a reason for thinking we should drop it. So we need reasons to think $(p \wedge \neg q)$ isn't always sufficient for the falsity of the indicative $(\mathit{if}\ p)(q)$. And there are (some say) concrete examples to be found. McGee's (1985) example:

(24) If a Republican won, then if Reagan didn't win Anderson did.

It seems possible to be in favor of a right-nested indicative like (24), and to be in favor of its antecedent, without being particularly attracted to the embedded conditional considered on its own. That doesn't yet fit the bill for what needs doing, but it is a start.[33] It is safe to say that this has not turned out to be the favored route.

Another option: perhaps indicative conditionals do not support import/export. As we have already seen, variably strict conditionals do not. So if indicatives are variably strict, then they won't. Stalnaker (1975) – rightly, I think – takes counterfactuals and indicatives to share the same core semantics. Since there are good reasons for the variably strict picture of counterfactuals, there are good reasons for that picture here, too. The reasons (as we'll see) have little to do with import/export.

Indicatives and counterfactuals are different: witness Adams pairs like (3a) and (3b). The idea is that they nevertheless share the common semantic core of variably strictness. What sets indicatives apart is a pragmatic constraint on their proper use: indicative conditionals say something about the not-yet-ruled-out worlds in a context.

Pragmatic constraint If p is compatible with the context c, then the $\leq_w$-minimal p-world(s) are compatible with c.

The constraint can be suspended – that, arguably, is what the distinctive tense/aspect marking of "subjunctive" conditionals does.

Strict conditionals are stronger than material conditionals, so we know that (under plausible assumptions about entailment) *or*-to-*if* inferences like (23) can't be entailments.[34]

Fact 15. If $(\mathit{if}\,\cdot)(\cdot)$ is a variably strict conditional then $p \vee q \not\models (\mathit{if}\,\neg p)(q)$.

Why then do instances of this pattern like (25) seem like entailments?

(25) Either the butler did it or the gardener did it.
So: if the butler didn't do it, then the gardener did it.

Stalnaker's answer: because there is a nearby pragmatic property – *reasonable inference* – that they do have.

Definition 12 (Reasonable inference). Suppose p is successfully asserted (at w in c) and c' is the resulting posterior context. $p, so: q$ is a *reasonable inference* iff q is accepted in c'.

Suppose that in order for a disjunction $p \vee q$ to be appropriately asserted in a context c that c has to have some $(p \wedge \neg q)$-worlds compatible with it and also some $(\neg p \wedge q)$-worlds compatible with it.[35] To be accepted in a context a sentence must be true at every world compatible with it. So if a disjunction is successfully asserted in c then $\neg(p \wedge q)$ is accepted in c'. Then anytime $p \vee q$ is appropriately asserted we will end up in a context in which the indicative $(\mathit{if}\,\neg p)(q)$ is accepted.[36]

Fact 16. If $(\mathit{if}\,\cdot)(\cdot)$ is variably strict and obeys the pragmatic constraint on indicatives then *or*-to-*if* is a reasonable inference.

This is an elegant solution. A crucial part of it is how it uses an independent and plausible pragmatic fact about disjunctions – that they are felicitous only when either conjunct might be true without the other – to get where we need to go. Therein lies the elegance. But there is no similar thing to be said about either right-nested conditionals like $(\textit{if}\ p)((\textit{if}\ q)(r))$ or their exported counterparts like $(\textit{if}\ p \wedge q)(r)$. So the strategy here leaves unexplained why import/export seems valid.

Collapse arguments rely (sometimes tacitly, sometimes not) on the assumption that indicative conditionals fit in with the rest of our linguistic tools in being vehicles that can be either true or false and have the same sort of semantic values that non-conditional language has. For that reason collapse arguments can be taken to be arguments for the thesis that indicative conditionals do not, after all, have truth-values or express normal semantic values. This is sometimes called the *no-truth-value* (NTV) theory.[37] The idea is that indicatives do not represent or report conditional information at all. Instead they are vehicles for expressing states of mind, in particular that the speaker has sufficiently high credence in the consequent given the antecedent. That is: they aren't things we assert, but things we use to make *conditional* assertions.[38]

There is a lot to be said for tying the appropriate use of conditionals to conditional assertion.[39] But NTV views are also hard to reconcile with a uniform picture of semantics. And, indeed, they are hard to square with why it is that conditionals mix with other, non-conditional, bits of language: they can be negated, they embed modal vocabulary, and they go in for quite a bit of nesting. Take the modals.

(26) a. If Red isn't in the box, Blue must be.
b. If Red isn't in the box, Blue might be.

Hypothesis: epistemic modality and indicative conditionals are tightly linked. Whatever kind of semantic-plus-pragmatic bundle we opt for in one case we should opt for the same sort of bundle in the other case. The dilemma in the offing here is that NTV either extends the thesis to the modals or not. If so, then since *must* and *might* aren't for asserting on their own, embedding them can't be something we do to conditionally assert them. So (26) are *prima facie* out. If not, then the attractive hypothesis linking modals and conditionals has to go.[40]

Even so, we may be driven in the NTV direction all the same. One reason (we won't really be able to do justice to it here): Lewis (1976) showed that (on pain of triviality) there is no two-place propositional operator such that the probability of the resulting proposition being true when given arguments p and q always equals $\Pr(q|p)$. Hence:

Fact 17. If $(\textit{if}\ p)(q)$ expresses a semantic value that determines a truth-value then it is not in general true that $\Pr((\textit{if}\ p)(q)) = \Pr(q|p)$.

In so far as we want our attitudes toward indicatives – the degree to which we believe them or want to assert them or bet on them or whatever – to track the corresponding conditional probabilities (and we do), that is an argument against assigning them truth-values.[41] But the argument can be resisted. As we will see, some theories (notably the "restrictor" theory developed by Kratzer, but also some others) seem to take triviality in stride.[42]

Another reason, due to Edgington (2008, §2; 2014): if indicatives have truth-conditions then those truth-conditions are either truth-functional or non-truth-functional. Neither seems right for indicatives. Consider non-truth-functional truth-conditions. (Truth-functional truth-conditions are a dead end because the horseshoe is the only

option.) Lack of truth-functonality requires *variability*. That is: the truth-values of p and q at w (in c) don't fix the truth-value of $(if\ p)(q)$ at w (in c). So if p is false at w and q is true at w then sometimes $(if\ p)(q)$ is true at w and sometimes it isn't. But indicatives seem to validate *or*-to-*if* and this requires *uniformity*. That is: if all the information you have is that $\neg p \vee q$ then that is always sufficient for $(if\ p)(q)$. Edgington says variability and uniformity are at odds and so no non-truth-functional theory can get this right. Conclusion: no way of assigning truth-conditions to indicatives can be right.

This is a compelling argument but it isn't airtight since there is room for a non-truth-functional theory to embrace both variability and uniformity. It is possible that the variability required by non-truth-functionality be variability in truth-value at a given world between contexts. And it is possible that the uniformity required by *or*-to-*if* be uniformity across worlds compatible with a given context. The master argument overlooks the possibility of such theories of indicatives.[43]

7 Antecedents as Restrictors

So far we have only been worrying about what various sorts of conditionals mean. There is an orthogonal issue: how are those meanings achieved by the *if*-of-English? We have assumed a straightforward answer: that the *if*-of-English is adequately represented in our regimented language by a binary connective $(if\ \cdot)(\cdot)$ and that this is assigned a uniform semantic value. But this is an assumption and can be challenged. In philosophy *can be challenged* often enough implies *has been challenged*. This is such a case.

Lewis (1975) argued that *if*s – apparently conditional expressions – in certain quantificational environments do not express any conditional thing. Example:

(27) $\left\{\begin{array}{c}\text{Always}\\ \text{Sometimes}\\ \text{Seldom}\end{array}\right\}$ if a farmer owns a donkey, he beats it.

What single thing could *if* mean in each of these and could that single thing be a conditional operator rightly so-called? The *if* might plausibly contribute some iffy meaning – for instance a strict conditional of some sort– when the quantifier is a universal like *always*. But conjunction is a better candidate if it's an existential like *sometimes*. Neither looks good for *seldom*.

In asking the question we assumed that we have two operators, a quantifier $@$ and a conditional $(if\ \cdot)(\cdot)$ and that the relevant structure of (27) sorts the scopes out this way:

(28) $@(if\ p)(q)$

That gets us into trouble so Lewis's (1975) conclusion is that sentences like those in (27) are not instances of a conditional operator plus an adverb of quantification.[44] We don't, appearances aside, have the two operators $@$ and $(if\ \cdot)(\cdot)$. Instead, he said, the *if*s in environments like these are a *non-connective* whose only job is to mark an argument slot for the adverb of quantification.[45]

The trouble-making feature is that *if*s under adverbs of quantification express restricted quantificational claims. In (27) those are claims about donkey-owning-farmer situations: that they are always/sometimes/seldom farmer-beating-donkey situations. There seems to be no way for the *if*s to both contribute a uniform conditional meaning and do this restricting. Since these *if*s restrict, they aren't iffy.

Kratzer's idea is that this holds not just for *if*s under adverbs of quantification but for all *if*s.[46] The picture is that *if* doesn't contribute a meaning, much less a conditional meaning, to the sentences in which it occurs:

> The history of the conditional is the history of a syntactic mistake. There is no two-place "if … then" connective in the logical forms for natural languages. "If"-clauses are devices for restricting the domains of various operators. (Kratzer, 1986, p. 11)

The thing all *if*s do is restrict operators (quantifiers, modals). So they aren't iffy. (What about bare conditionals, conditionals which appear not to have a nearby operator? Since the hypothesis is that *if*s are devices for restricting operators, there must be one. So we posit a covert, universal operator.)

Definition 13 (Restrictor analysis). Conditionals are restricted operators with logical forms like this:

$$\text{Quantifier/operator + if-clause + consequent clause} \quad \textsc{operator}_{(p)(q)}$$

Two things to notice. First: the binary connective $(\textit{if}\ \cdot)(\cdot)$ makes no appearance. The action is all about the operator. Second: for just that reason we won't have a story about "indicative conditionals" or "counterfactual conditionals." Instead, as we'll see, we get a story by saying something about the operators that the relevant *if*-clauses restrict. That's why this is less an issue about *what* various conditionals mean and more an issue about *how* natural language sentences manage to carry those meanings.

Figure 17.3 sketches the differences in structure. The left and center trees illustrate the structure where an operator and a conditional connective mix: the left is the case where the operator outscopes the conditional with antecedent p and consequent q and the center is the case of a conditional with antecedent p and complex consequent OPERATOR q. The rightmost tree shows the structure for the restrictor analysis. Here the "antecedent" is the restrictor for OPERATOR and the "consequent" q is the nuclear scope. (We'll now focus on the case where the operators involved are modals (as opposed to quantifiers).)

Take an ordinary counterfactual like (29):

(29) If Alex had been there, she would have had fun.

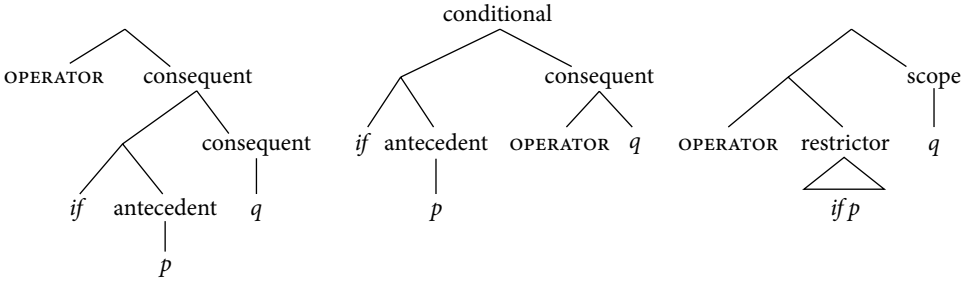

Figure 17.3 Operators + *if*: conditional versus restrictor.

So far, treating the (counterfactual) *if*-clause as a restrictor doesn't do much because we haven't yet said what the modal *would have* means.

In principle we could adapt just about any theory about what such conditionals mean to what such modals mean. Here we will (with some simplifications) follow Kratzer and give an analysis that makes the same predictions as the variably strict story. Here is the idea: modals quantify over a domain, saying that all or some of the best worlds in it (and in which some additional thing (provided by a restricting *if*-clause if there is one) is true) satisfy the prejacent (the scope). Context supplies the domain (the *modal base* at a world) and the notion of bestness (induced by the *ordering source* at a world).

Definition 14 (Modal base, ordering source). A *conversational background* is a function from worlds to sets of propositions. A *modal base B* is a conversational background. Similarly, an *ordering source O* is a conversational background. The *ideal worlds* at w relative to B and O is the set of worlds that are $\leq^{O}_{w}$-minimal in $\cap B(w)$ (where $\leq^{O}_{w}$ is the ordering induced by the set of propositions $O(w)$ used in Fact 10).

Contexts determine both a modal base and an ordering source. These are the same type of object, but they play different roles. Modal bases simply form domains. We aren't making use of their additional structure here, so we just take the image of a base B_w to get the worlds compatible with all the propositions in it. The extra structure in ordering sources – the extra structure over and above a set of worlds – is used. We use it to induce an ordering. The two pieces together get us the ideals.[47]

Context has the job of selecting values for the modal base and ordering source.[48] Once it does, the modal simply contributes (restricted) quantificational oomph.

Definition 15 (Modals). Suppose B is the modal base and O the ordering source determined by context c. Then:

(i) $[\![\textit{must/would}(p)(q)]\!]\, c,w = 1$ iff for every ideal v at w wrt $B \cup \{[\![p]\!]\}$ and O: $[\![q]\!]\ c,v = 1$.

(ii) $[\![\textit{might/might have}\ (p)(q)]\!]\, c,w = 1$ iff for some ideal v at w wrt $B \cup \{[\![p]\!]\}$ and O: $[\![q]\!]\ c,v = 1$.

The restrictor p simply augments the modal base. And so, together with Definition 13, that's all that *if*-clauses do.[49]

We have barely constrained what sorts of functions the backgrounds can be. Adding constraints gets us different possibilities for the flavor of the modal involved and the sort (and strength) of conditional. Some possibilities for a conversational background f:

- f is *realistic* iff for every w: $w \in \cap f(w)$.
- f is *totally realistic* iff for every w: $\cap f(w) = \{w\}$.

Backgrounds, like accessibility relations and accessibility functions, can also characterize sorts of contextually relevant information:

- f is an *epistemic* background iff (for every w) $f(w)$ is the set of propositions known to the c-relevant agents at w.

- *f* is a *stereotypical* background iff (for every *w*) *f*(*w*) is the set of propositions which normally hold at *w*.

And so on.

This predicts for a wide range of possible contextual resolutions for conditionals.

Fact 18. Material conditionals correspond to the special case where *B* is totally realistic and *O* is empty. Strict implication corresponds to the special case where both *B* is empty and *O* is empty.

To get a feel for this, let's walk through the material conditional case. Let *B* be a totally realistic background and *O* be empty. Then a conditional *if p*, (*must*) *q* – that is, *must* (*p*) (*q*) – at *w* says that the best worlds in $\cap B(w)$ that are *p*-worlds are all *q*-worlds. But $\cap B(w) = \{w\}$ and the induced ordering treats all worlds alike. So whether the conditional is true is down to *w* alone. Suppose *p* is true at *w*. Then the conditional is true iff *w* is a *q*-world. Suppose *p* is false at *w*. Then the conditional is vacuously true at *w* since there are no *p*-worlds in $\cap B(w)$. These are just the truth-conditions of the material conditional.

The empirical hypothesis is that counterfactuals are conditionals (which is to say "conditionals") interpreted against an empty modal base and a totally realistic ordering source. This is backed up by the following:

Fact 19. Let *B* be the empty modal base and *O* be the *c*-relevant totally realistic ordering source. Now take the premise set *P* to be such that for every *w*: $P_w = O(w)$. Then: $[\![\text{would}(p)(q)]\!]\ c,w = 1$ according to the restrictor analysis (Definition 15) iff $[\![(\textit{if}\ p)\ (q)]\!]\ c,w = 1$ according to the premise set analysis (Definition 7).

So premise sets correspond to special ordering sources and the analyses use them in the same way. Hence the (promised, and now delivered) equivalence.

All of this carries over to the case of indicatives. There the relevant modals are epistemic, taking an epistemic modal base and an empty ordering source. Doing things that way, as Kratzer (1986) argues, relieves some of the pressure in the collapse arguments we saw in §3 and §6. I want to mention just two examples for now.

The first: import/export. The trouble was that we couldn't, without collapse to the material conditional, both satisfy this and have the truth of a counter-example ($p \wedge \neg q$) always suffice for the falsity of the corresponding indicative (*if p*)(*q*). (Or, in this case, the corresponding modal *must*(*p*)(*q*).) Epistemic modal bases are realistic (since you can't know false things), so it would seem that we can't have import/export. But there is room for maneuvering for the restrictor view.

Fact 20. *must* (*p*) (*must* (*q*) (*r*)) $\dashv\models$ *must* (*p*$\wedge$*q*) (*r*)

How is this possible? Kratzer says that Gibbard's proof "does not threaten" this analysis, but doesn't really say more than that. The answer lies in the fact that conditionals ("conditionals") express different things when they occur on their own compared to when they occur embedded.[50]

The second: it is hard to see how the restricting behavior gets done if *if* expresses a *bona fide* conditional. This is part of what is going on in Lewis-style-triviality arguments. But there is an easy way out.

(30) a. If the bet is on odd, it's probably a loser.
probably $(p)(q)$
b. If the coin is fair, then the probability of heads is $\frac{1}{2}$.
$\frac{1}{2}$-*probable* $(p)(q)$

Filling in very off-the-shelf semantics for the probability modals here immediately predicts that (30a) is true iff the corresponding (c-relevant) conditional probability is high enough and that (30b)is true iff the relevant (c-relevant) conditional probability is 0.5. Triviality doesn't threaten because there is no separate conditional proposition that the probability modal takes scope over.[51]

Again, in principle the *if*s-as-restrictors picture is compatible with just about any story of the relevant modals and so can be made to come out equivalent to just about any story of conditionals. To reiterate: that fact goes in the pro column since one of the achievements here is to separate out questions about how conditional meanings get expressed from questions about what those conditional meanings are.

8 Dynamics and Indicative Conditionals

The dynamics of counterfactual domains highlights ways in which conditionals can be seen as *externally* dynamic: interpreting one sentence changes an essential component of conversational score that is relevant for the interpretation of a later sentence. Conditionals can also be thought of as *internally* dynamic: interpretation of one part of a sentence changes an essential component of conversational score that is relevant for the interpretation of a later bit of that same sentence.[52]

Let's explore that by sketching a dynamic semantics of indicative conditionals that is inspired by the Ramsey test. In fact, we will see two implementations of that theory: one implementation in a dynamic framework and one using the (seemingly) more familiar context-plus-index-delivers-truth-values setup we've mostly been assuming throughout. The first implementation is in Gillies (2004) and the second in Gillies (2009; 2010).[53] Taking indicatives to be dynamic (strict) conditionals offers a new perspective on the various collapse arguments, Edgington's "master argument" against truth-conditions, and a modal parallel to Lewis's argument about *if*s under adverbs of quantification.[54]

We have been assuming throughout that semantic values are sets of points of evaluation: they are the kind of thing that, when supplied a world, deliver a truth-value (that's what Figure 17.1 illustrates). Nothing prevents us from using this setup to understand more complicated object languages than we have so far. For instance, we might have an object language that includes recipes and programs.[55] A regular (modal) sentence has a set of worlds as its semantic value because what it says is that things are a certain way. Programs aren't like that. They instead say what they do to ways the world might be: execute it in a world like this, such-and-such results; execute it in a world like that, thus-and-so results. So the denotation of a program is a relation, a set of ordered pairs of worlds (or "states" or whatever the points of evaluation are): the set of ordered pairs such that executing the program in the first member of the pair (possibly) terminates in the second member of

the pair. Dynamic semantic theories can be thought of as theories that say *all* sentences are like that: they are recipes or programs for changing the context or conversational score.[56]

So much for broad brushstrokes. In order to make this concrete we need to decide what aspects of conversational score we care about or what sort of information we want represented in a context. (We will do this in an update semantic framework (à la Veltman, 1996) and follow custom by sometimes calling such bodies of information *information states*.) For us this decision is easy: sets of not-yet-ruled-out possibilities.

Definition 16 (Contexts, information states). A context or information state s is a subset of the set W of possible worlds. I is the set of all such s's.

The limit cases of contexts I (the state of total ignorance) and $\varnothing$ (the absurd state) will have roles to play.

Sentences of our intermediate formal language (that is, propositional logic plus $(\mathit{if}\,\cdot)(\cdot)$) have the same type of semantic value as programs. They are *context change potentials*, relations between information states (actually: functions from states to states).[57]

Definition 17 (Context change potentials (CCPs)). Let a be any atomic formula, p, q be any formulas, and s be any information state. Then the semantic value function $[\cdot]$ is defined as follows:

(i) $s[a] = \{w \in s\colon\ w(a) = 1\}$
(ii) $s[\neg p] = s \setminus s[p]$
(iii) $s[p \wedge q] = s[p][q]$
(iv) $s[(\mathit{if}\,p)(q)] = \{w \in s\colon\ s[p][q] = s[p]\}$

Glossing the machinery for the first three clauses: (i) the change induced by a successful assertion of an atomic sentence is to remove possibilities in which the atom isn't true; (ii) negation corresponds to complementation; and (iii) conjunction is functional composition (the state $s[p]$ is the input for the function $[q]$). All of these represent programs whose purpose is to change the conversational score. The indicative conditional is a different sort of program. It is a *test*: it checks to see if the state $s[p]$ is the same as the state $s[p][q]$. If so, the result is the original s; otherwise, the test fails and the result is $\varnothing$.

This is represented in Figure 17.4 where CCPs are transitions between states. Updating s with $[(\mathit{if}\,p)(q)]$ loops to s iff updating $s[p]$ with $[q]$ loops to $s[p]$. That is the situation in the

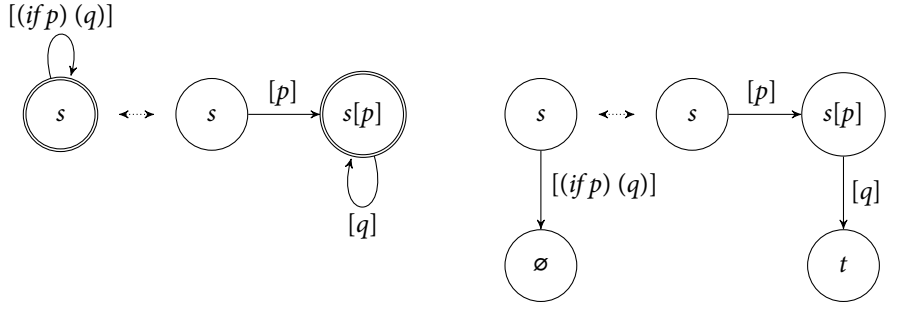

Figure 17.4 Conditional updates: accepted (left) and not (right).

left diagram. On the right: $s[(\mathit{if}\ p)(q)] = \varnothing$ because updating $s[p]$ with $[q]$ adds something new, carrying us to new state t.

This represents a conservative extension of the ordinary semantics of propositional logic.

Fact 21. Let p be any *if*-free formula and s any state. Then: (i) $W[p] = [\![p]\!]$ and (ii) $s[p] = s \cap [\![p]\!]$.

The indicative, though, makes a difference. The way this is often demonstrated is to note that the update function $[\cdot]$ is not, in general, distributive in the sense that it doesn't respect arbitrary unions of its arguments.

Fact 22. It is not in general the case that $s[p] = \bigcup_{w \in s} \{w\}[p]$.

Negated conditionals are a counter-example: updating a state $\{w, v\}$ where $w(a) = w(b) = v(a) = 1$ and $v(b) = 0$ with $\neg(\mathit{if}\ a)(b)$ returns $\{w, v\}$ but updating the singletons $\{w\}$ and $\{v\}$ and collecting the results returns just $\{w\}$.

So indicatives are strict conditionals that invite tests. But when are they true? And what about entailment? Since we are in a richer logical space, we have options. These are reasonable choices: truth is a fixed-point, entailment is so-called *update-to-test* entailment.[58]

Definition 18 (Truth, entailment). For any sentences p, p_i, q and any state s:

(i) p is *true* in s (or s *supports* p) iff $s[p] = s$.
(ii) $p_1, \ldots, p_n \models q$ iff for any s: q is true in $s[p_1] \ldots [p_n]$.

Together with Fact 21, it follows that (for *if*-free p) $s \models p$ iff $s \subseteq [\![p]\!]$. This picture of indicatives is Ramseyan in that the truth of an indicative in a state rides on whether a derived or subordinate sub-state supports the consequent.

Fact 23. An indicative $(\mathit{if}\ p)(q)$ is true in s iff q is true in $s[p]$.

This too can be seen in Figure 17.4 where double-circled nodes represent supporting states.

Collapse arguments highlight the puzzle indicatives pose: they seem to say more than material conditionals but there are no good candidates for what that more is (or worse: it seems impossible for them to say more). The dynamic turn opens up possibilities.

Fact 24. The dynamic conditional together with update-to-test entailment supports *modus ponens* and import/export. Moreover, if p implies q then $(\mathit{if}\ p)(q)$ is true in any state and the $(\mathit{if}\ \cdot)(\cdot)$ is stronger than the corresponding material conditional.

This was a package that collapse arguments led us to believe was not possible.

How does this square with the fact that *modus ponens* plus import/export lead to collapse (Fact 5)? The frameworks are different and so comparisons non-obvious. Let's focus on the conditional-only fragment of the language and (following Gillies, 2009; 2010) retell things

in the context-plus-index-delivers-truth-values setup.

As in §3, contexts deliver sets of possibilities relevant for indicatives. We will again take contexts themselves to be such functions. Here we choose two properties such functions must have.

Definition 19 (Well-behaved contexts). c is well behaved iff for any w:

(i) c is reflexive: $w \in c(w)$;
(ii) c is euclidean: if $v \in c(w)$ then $c(w) \subseteq c(v)$.

The first means that the facts at w are always relevant to the truth of indicative at w. The second means that what is relevant at w is settled. Reflexive and euclidean functions are closed: if $v \in c(w)$ then $c(w) = c(v)$.

The theory here will be a strict conditional theory (different, as we'll see, from the one in Definition 2). So it will say that $(\textit{if}\ p)(q)$ is true at w in c just in case the p-worlds in $c(w)$ are all worlds at which q is true. But true at what context? The Ramseyan subordinate or derived context – write it $c + p$ – got by hypothetically adding the information of the antecedent to c.

Definition 20 (Shifty strict conditional). For any p, q in a well-behaved context c:

$$[\![(\textit{if}\ p)\ (q)]\!]\, c,w = 1 \text{ iff if } v \in c(w) \text{ and } [\![p]\!]\, c,v = 1 \text{ then } [\![q]\!]\ c{+}p,v = 1$$

where $c + p = \lambda w\,.\, c(w) \cap [\![p]\!]c$.

This replicates the dynamic strict conditional.[59] And so there is no collapse. Notice that this puts conditional antecedents to work twice over. They restrict the domain throughout which we check for the consequent's truth. Plain (non-shifty) strict conditionals do this, too. But here antecedents also contribute to the context that is relevant to the interpretation of those consequents. This is the internal dynamics we saw earlier and sets this apart. The shifting is essential. One way to see that: strict conditionals with no shifting can't validate import/export. Another way: earlier we saw that Edgington's master argument against truth-conditions seemed to pit variability against uniformity but that the argument overlooked the possibility that the variability is variability in truth-value at a world across contexts and the uniformity is uniformity across worlds within a context. The context-shifting delivers exactly that. And a third way: we saw that from the point of view of correspondence theory, strict conditionals that support import/export have peculiar selection functions (Fact 4). Not so here: with a context-shifty strict conditional the odd property of "shift coreflexivity" that characterizes one direction of import/export gets replaced by the much more tame property of transitivity.[60]

Let's wrap up this brief foray into the shifty/dynamic strict conditional by looking at one more way the dual roles assigned to antecedents make a difference. The problem that (27) exposes for *if*s that mingle with adverbs of quantification has parallels for *if*s that mingle with (epistemic) modals.

(31) a. If Red isn't in the box, Blue must be.
b. If Blue isn't in the box, Red must be.
c. If there is a marble in the box, it might be Red.

Assume that *must* here expresses a strong necessity epistemic modal $\Box$ over the set of relevant worlds.[61] And assume that *might* is its dual. Then it is surprisingly difficult to say that

if contributes the same iffy meaning to all the examples in (31).

Here's part of the trouble.

(32) Red might be in the box and Blue might be in the box.

This is perfectly compatible with (31a) and (31b). For instance: suppose I know that one and only one of my marbles – Red, Blue – is in the box. But for nearly every p and q the conditionals $(if\ \neg p)(\Box q)$ and $(if\ \neg q)(\Box p)$ aren't jointly compatible with $\Diamond p \wedge \Diamond q$. The restrictor theory has no trouble here.[62]

The broader issue here that parallels what Lewis saw with adverbial quantifiers is that *if*-clauses restrict these modals. But it is hard to see how that restricting can get done by a single conditional operator: one sort of operator might seem reasonable if the operator is a universal like *must* but that same operator looks pretty bad for an existential commingling modal like *might*.[63] This is an issue that shifty/dynamic conditionals handle with ease.

Fact 25. Given Definition 19 and Definition 20:

(i) $(if\ \neg p)(\Box q), (if\ \neg q)(\Box p)$ are jointly compatible with $\Diamond p \wedge \Diamond q$
(ii) $(if\ p)(\Diamond q) \dashv\models \Diamond (p \wedge q)$

These are the same predictions that the restrictor analysis achieves but here *if* still makes a uniform contribution. That is because of the Ramseyan core: the antecedent contributes to the subordinate or derived context that the embedded modal is sensitive to.

9 Other Surveys

No survey of conditionals can be both complete and digestible. This one aimed at the latter property at the expense of the former property and so strategic choices dictated that things which deserve more attention did not get their due. Other surveys – for instance, Bennett (2003), Edgington (1995; 2008), von Fintel (2012a; 2012b), Gillies (2012), Kaufmann and Kaufmann (2015) – make different choices and thus emphasize different parts of the terrain. There are, in addition, more substantial investigations of the logic of conditionals, including Nute (1980) and (especially) Part II of Veltman (1985). A healthy preoccupation with *if*s should be fed a diet rich in just this kind of variety.

Notes

1 Counterfactuals are sometimes called "subjunctive conditionals," but this is an even worse fit. A natural thought: the "subjunctive" marking goes exactly with counterfactuality. Alas, no, as an example from Anderson (1951) shows:

(33) If Jones had taken arsenic, he would have shown just exactly those symptoms which he does in fact show.

A doctor may well use a conditional like this to argue that what afflicts poor Jones is arsenic

poisoning. If counterfactuality means anything like (the speaker is taking it that) the antecedent is false, then the distinctive marking isn't sufficient counterfactuality. Nor is it necessary. In sportscasterese, seemingly run-of-the-mill indicative conditionals are used to convey counterfactual meanings. It's the top of the ninth, the visiting team is down a run but they are down to their last out with a runner – Speedy, as it happens – on second. Slugger hits a double to the gap – surely Speedy should score! – but as Speedy rounds third he trips, falls, and is thrown out. The visitors lose in a shocker. The announcer almost can't believe it. In the aftermath he says:

(34) If Speedy stays on his feet, they probably win the game.

The announcer isn't confused about how the game ended: he's sure Speedy would have scored and Slugger would likely have been hit in, too. See von Fintel 1998 (and the references therein) for the status of the connection between "subjunctive" marking and counterfactuality.

2 There are complications and wrinkles galore. Here are just three, followed by executive decisions about the issues they raise. One: some conditional constructions do not seem to carry conditional information at all.

(35) There are biscuits on the sideboard if you want them. (Austin, 1956)

These are so-called *biscuit* (or relevance) conditionals and are fine things to say but they don't normally (or obviously) express what conditionals normally (or obviously) do. So let's agree to set them aside. Two: there are ways to conventionally express conditional information in natural language without resorting to *if*. This is especially clear with conditional imperatives:

(36) a. Be on time or text me!
b. Keep it up and I'll turn this car around!

These do manage to express some conditional kind of meaning (iffiness but *if* lessness!). Still, set them aside. (While we're at it: we will also not have anything to say about conditional imperatives or conditional questions.) And three: Some languages (apparently) lack lexicalized *if*-constructions altogether. We will largely ignore cross-linguistic pressures, too. The hope is that the executive decisions won't distort things (too much).

3 Though this is a survey, it can't (and won't) pretend to be either exhaustive or unopinionated. See §9 for references that may provide balance on either horn.

4 Whether conditionals in natural language can be represented by a binary conditional connective in a regimented intermediate language is also, as we'll see, up for grabs.

5 Lewis (1980) argued that contexts and indices are both needed and that neither can do the work of the other.

6 For now. Later, when we flirt with various dynamic theories, all we will need to know about them is that they are the kind of thing that *atomic* sentences (of our intermediate language) are true or false at. If an atomic a is true (false) at w, we'll say that $w(a) = 1$ ($w(a) = 0$).

7 So-called because the material conditional is sometimes symbolized by the horseshoe $\supset$.

8 This is Grice's (1975) strategy and is taken up and extended by Lewis (1976). Another pragmatic defense, developed by Jackson (1991), says that when you use an indicative $(if\ p)(q)$, it conventionally implicates that your credence in its truth (that is, the truth of $p \supset q$) conditional on p is high. (Lewis, in the "Postscript" to his 1976, drops the conversational story and instead goes for a slight variant of Jackson's.) There is little independent evidence in favor of this stipulated conventional implicature.

9 *All the beer is gone!* often does not mean that the universe is out of beer but something more modest like the relevant beer supply is out.

10 This discussion follows the discussion in Kaufmann and Kaufmann (2015) and to some extent some of the discussion in van Benthem (2001, §3.2). As we'll see, this provides a modal perspective to Gibbard's (1981) argument that reduces any "propositional" operator meeting a few

minimal assumptions to the material conditional.

11 To see that if $v \in c(w)$ then $v \in c(v)$ implies the right-to-left direction of import/export: assume that $(if\ p)((if\ q)(r))$ is true at w wrt c and consider any $v \in c(w)$ such that $[\![p \wedge q]\!] c,v=1$. (We want to show that $[\![r]\!]\ c,v = 1$ and hence that $(if\ p \wedge q)(r)$ must be true at w wrt c.) Since $[\![(if\ p)((if\ q)(r))]\!] c,w = 1$, it follows that all p-worlds in $c(w)$ are worlds where $(if\ q)(r)$ is true. Hence since $[\![p]\!]\ c,v = 1$, it follows that $[\![(if\ q)(r)]\!]\ c,v = 1$. So all q-worlds in $c(v)$ must be worlds where r is true. Since if $v \in c(w)$ then $v \in c(v)$ and $[\![q]\!]\ c,v = 1$, it thus follows that $[\![r]\!]\ c,v = 1$, as desired. (For the other direction(s): it's generally easier to go by showing how if c isn't constrained (e.g., if it's not in general true that $v \in c(w)$ implies $v \in c(v)$) then you can use the witnessing worlds to counter-example the relevant logical principle. Providing full proofs of these correspondences is left as a rainy-day exercise for the reader.)

12 That is:

(i) *All As are Cs* $\vDash$ *All Bs are Cs* whenever $B \subseteq A$;
(ii) *All As are Bs*, *All Bs are Cs* $\vDash$ *All As are Cs*;
(iii) *All As are Bs* $\vDash$ *All non-Bs are non-As*.

13 Lewis credits Sobel for pointing them out (Sobel, 1970).

14 To be sure: the spatial metaphor here is, in fact, a metaphor. We'll return to this at the end of this section.

15 Sort of. We might go instead with *weak centering* instead of centering: $w \leq_w v$ for every v. This requires that w is always among the closest worlds to w, but allows that other worlds might be equally close. There aren't decisive reasons for favoring weak centering over centering since both deliver *modus ponens* for the counterfactual. Or we might (as Pollock, 1976, does) weaken the ordering assumption by permitting incomparabilities as well as ties in the ordering.

16 Lewis offered this reason against the limit assumption. Take a counterfactual like this:

(37) If Alex had been a little taller, she would have played basketball.

No matter Alex's height, there is no world *closest* to ours in which Alex is just a little taller. Take any world v where Alex's height is h_v where that is a bit taller than her height $h_@$ here. There is another world u such that her height h_u is intermediate. That's just how height works. The argument is not quite decisive, though, since giving up the limit assumption means giving up some pretty desirable features, too. (See, especially, Pollock, 1976; Herzberger, 1979.) And see Swanson (2012) for a discussion of preserving conditional excluded middle without the Limit Assumption.

17 Just what kind of signal? There is a debate. It won't matter too much for us, so set it aside. But see Stalnaker (1975), Veltman (1976), and von Fintel (1998).

18 That is: rendering a conditional determinately true iff if it's true according to each admissible order. Stalnaker (1984) develops this approach in his defense of conditional excluded middle.

19 The famous footnote: "If two people are arguing 'If p, will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q…" (Ramsey, 1929/1990, p. 155). The adding this involves can't be feigned or hypothetical belief.

(38) If my students are cheating in class, then I will not discover it (because they're so clever).

The conditional is (let's say) true even though the state I get into by feigning belief in the antecedent won't be one in which I believe the consequent. (Thomason is credited with this observation in van Fraassen, 1980.) The right way to understand augmenting for the Ramsey test is restricting a body of information (my belief state or the contextually relevant information or whatever) by the content of the antecedent. That's different.

20 The (independently developed) classics: Kratzer (1981b) and Veltman (1976), developed and extended in (respectively) Kratzer (1989) and Veltman (2005).

21 A subset P^*_w (also a premise set at w!) of P_w is X-consistent iff $\cap P^*_w \cap X \neq \varnothing$. Such a subset is *maxi-*

mal iff it is contained in no other such subset of P_w.

22 There is nothing in the premise semantics setup that mentions nearness or similarity or anything like that. So to argue that the nearness metaphor in the Lewis–Stalnaker setup has to be more than a metaphor we need an argument that the corresponding premise sets somehow encode (by what is lumped together and what isn't) nearness or similarity information. I don't know what that would look like.

23 One sort of challenge that we won't discuss is whether there is any substantive notion of similarity that plays the role that some versions of the variably strict theory carve out for it. There are different threads one can tug here. One begins with Fine (1975) and (what can be read as) Lewis's (1979a) reply. This thread is neatly discussed in Bennett (2003). Another thread: similarity is the wrong sort of thing. This argument comes from the premise semantics corner (Tichy, 1976; Kratzer, 1989; Veltman, 2005). And a third thread challenges that there can be no such thing as all-in similarity because there is no sensible way of aggregating the comparisons that the relevant theories say must be aggregated (Morreau, 2010).

24 von Fintel (2001) credits Irene Heim with the observation.

25 Lewis saw the loophole in his argument but didn't take it seriously:

> It is still open to say that counterfactuals are vague strict conditionals … and that the vagueness is resolved – the strictness is fixed – by very local context: the antecedent itself. That is not altogether wrong, but it is defeatist. It consigns to the wastebasket of contextually resolved vagueness something much more amenable to systematic analysis than most of the rest of the mess in that wastebasket. (Lewis, 1973, p. 13)

26 The version we will look at says that the meaning of counterfactuals can be factored into the context-changing part of meaning and the content-assigning part of meaning. The version in Gillies (2007) is dynamic all the way, based on an argument that (i) *might*-counterfactuals exhibit the same sort of context-affecting as *would*-counterfactuals with strengthened antecedents and (ii) the only way to accommodate that behavior is by treating the meaning of counterfactuals as their context change potentials.

27 There are a number of equivalent ways of implementing things; in the text, I opt for one that makes use of Lewis-style systems of spheres. A few small points: (i) even though officially what is counterfactually relevant can depend on the world at which we evaluate a counterfactual, since we are going to ignore iterated counterfactuals I'll often drop reference to the world; (ii) I'll treat the default situation as one where only the world of evaluation is counterfactually relevant (rather than a set including it); this is to make comparisons easy to see (for a setup that doesn't go this way see Gillies, 2007, §8); (iii) let's assume (as we have been sometimes doing already) that W is finite.

28 The classic references are Karttunen (1973), Stalnaker (1974), and Lewis (1979b).

29 There is room for refinement here if we want to address the (alleged) dynamics associated with *might*-counterfactuals (Gillies, 2007, §§7–8). That will take us too close to the bleeding edge for today.

30 We have put things in terms of systems of spheres. That was optional since there is an obvious back-and-forth between such a system $c(w)$ centered on w and an ordering $\leq_w$: $u \leq_w v$ iff u belongs to every sphere in $c(w)$ that v does (Lewis, 1973).

31 Note that $c|p|(p \wedge r)| = c|(p \wedge r)|$.

32 Should we deny that indicatives are bound from above by strict implication? I know of no good reason to pursue this.

33 We can't do justice to the literature surrounding McGee-style counter-examples (but see the references in Bennett, 2003, and the more recent discussions in Gillies, 2004; Weatherson, 2009; and Huitink, 2012). Along these lines: there has been some resurgent interest lately in generating "counter-examples" to various principles for conditionals. (See, for instance, Kolodny and MacFarlane, 2010.) I am of the opinion that this is largely a mischaracterization of the issues. For

example: there is no way to say what counts as *modus ponens* without saying what counts as "entailment." And where some might see *modus ponens* failures I'm more inclined to see a poor choice for a story about entailment. A less cryptic version of the point can be found in Gillies (2009).

34 Counter-example: suppose there are just two c-relevant worlds, w and v, where w is a $(\neg p \wedge q)$-world and v is a $(p \wedge \neg q)$-world. Then $p \supset q$ is true at w in c. Since p is compatible with c, the $\leq_w$-minimal p-world must also be compatible with c: so it's v. But since v is a $\neg q$-world, then $(\textit{if}\ p)(q)$ isn't true at w.

35 This runs into some trouble if either p or q is necessary. That's likely a feature of the possible worlds framework as much as a bug of the quasi-Gricean idea here about disjunction, so let's set it aside.

36 Argument: suppose $p \vee q$ is appropriately asserted at w in c. So c has both $(p \wedge \neg q)$-worlds and $(\neg p \wedge q)$-worlds compatible with it. Let c' be the posterior context; it has only $(p \wedge \neg q)$-worlds compatible with it. Let v be any world in c'. The indicative $(\textit{if}\ \neg p)(q)$ requires that the nearest $\neg p$-world(s) to v be compatible with c'. So they must be q-worlds.

37 See Adams (1975), Gibbard (1981), Bradley (2000), Edgington (2008); and Bennett (2003).

38 This is different from the versions of the horseshoe theory in Lewis (1976), Jackson (1991) and Lewis (1986) according to which indicatives have the truth-conditions of material conditionals but which, owing to some pragmatic wrangling, have high (enough) conditional credence as assertability conditions.

39 Stalnaker (2005) tries to narrow the gap between a conditional propositions picture and a conditional assertion picture.

40 There are things that are sometimes said on behalf of the NTV view here (see, for instance, Bennett, 2003) but I don't find them compelling. A more thorough discussion can be found in Gillies (2012).

41 The basic result has been extended and fortified many times over: see, for instance, Lewis (1986), Gärdenfors (1982), and Hájek (1989, 1994) and in the qualitative domain, Gärdenfors (1986) and Segerberg (1989).

42 For those in the know or those returning to this section after we discuss the restrictor theory in §7: the idea there is that when indicatives and probabilistic hedges mix, the correct logical form isn't one of a conditional scoping under (or over) a probability operator. Rather it's a dyadic probability operator whose first argument is the "antecedent" and whose nuclear scope is the "consequent." This idea is advanced in Kratzer (1986; 2012) and resisted in various ways in Rothschild (2012) and von Fintel and Gillies (2014).

43 Dynamic or so-called "context-shifty" accounts, as we'll see in §8, occupy just this spot in logical space. The master argument overlooks this possibility. The connection to Edgington's master argument is discussed in Gillies (2009, §4).

44 Saying that the scoping is $(\textit{if}\ p)(\mathcal{Q}q)$ fares no better.

45 Obligatory quote:

> The *if* of our restrictive *if*-clauses should not be regarded as a sentential connective. It has no meaning apart from the adverb it restricts. The *if* in *always if* …, …, *sometimes if* …, …, and the rest is on a par with the non-connective *and* in *between* … *and* …, with the non-connective *or* in *whether* … *or*, …, or with the non-connective *if* in *the probability that* … *if* …. It serves merely to mark an argument-place in a polyadic construction. (Lewis, 1975)

46 Developed (with some variations) in, e.g., Kratzer (1981; 1986; 1991; 2012).

47 As before, and for the same reasons, the Limit Assumption (or finiteness, take your pick) is in play here.

48 That's no different from the variably strict story where contexts were asked (somewhat implicitly) to supply an ordering of relative similarity $\leq_w$ or a set of premises P_w for each world w.

49 Two questions. What if there is no restrictor (that is, if the modal claim is something like *Alex*

must be there)? Then take the first argument to be your favorite tautology. What if there is no modal at all (that is, if the conditional is a bare conditional like *If Alex is at the party, she is outside*)? Since *if*s restrict, there must be one. So posit a covert necessity operator (a covert *must*).

50 See Gillies (2009) and Khoo (2013).

51 This is a very attractive way out, one hinted at (but somehow not pursued) by Lewis himself when he calls the *if* in *the conditional probability of … if …* a non-connective. It has, however, recently been challenged (Rothschild, 2012; von Fintel and Gillies, 2014).

52 For those familiar with dynamic predicate logic (Groenendijk and Stokhof, 1991): the existential quantifier is externally dynamic since its (semantic) binding power outruns its (syntactic) scope; conjunction, which is relational composition, is internally dynamic.

53 The original idea was inspired (in roughly equal parts) by Ramsey's footnote (1929/1990), the data semantic treatment of indicatives in Part III of Veltman (1985), and the simple update semantics for epistemic modals in Veltman (1996). (The inspiration and divergences, I trust, will be clear.) There is a point to having both implementations available, but we will get to that.

54 There are further refinements, developments, and purposes the dynamic perspective on conditionals can be put to that we won't be able to do justice here. See, for instance, Starr (2014) and Willer (2014).

55 That's what propositional dynamic logic, a particular modal logic, is. A good reference: Harel, Kozen, and Tiuryn (2000).

56 The theories I have most in mind here as exemplars of (classic) dynamic semantic theories are those in Groenendijk and Stokhof (1991) and Veltman (1996). Besides the analogy to the semantics of programming languages (and the empirical coverage they offer), there are other inspirations for the early dynamic turn, especially Karttunen (1973), Stalnaker (1974), and Lewis (1979b).

57 The notation in Definition 17 is post-fix: s is the argument to the function $[p]$.

58 Much of what we know about the landscape here – what options there are, what the properties of those options are – is due to Veltman (1996), van Benthem (1996), and van der Does, Groeneveld, and Veltman (1997).

59 Officially we need to define an entailment relation (Gillies, 2009, §6). The simple case: $p_1, p_2 \vDash q$ iff if $[\![p\ 1]\!]c,w = 1$ and $[\![p\ 2]\!]c{+}p\ 1,w = 1$ then $[\![q]\!]((c{+}p\ 1){+}p\ 2),w = 1$.

60 See Kaufmann and Kaufmann (2015).

61 Each part of this assumption is defensible – that *must* is strong (von Fintel and Gilles, 2010) and that it quantifies over the *if*-relevant worlds (Gillies, 2010, §6).

62 The reply, invariably, is that we should wide-scope the modals. This turns out to force a choice between giving up a link between *if* s and *must*s:

(39) If Alex is here, Billy is here ≈ If Alex is here, Billy must be here.

and giving up a link between *if* s and *might*s:

(40) If Alex is here, Billy might be here ≈ It might be that Alex and Billy are here.

The restrictor theory doesn't have to give up either. This is discussed in Gillies (2010, §6).

63 Lewis admitted that the restricting was after all possible with a "far-fetched" interpretation of the conditional – the one defended in Belnap (1970) that says that $(if\ p)(q)$ is true at w if the confirming instance is true, false if the counter-example is true, and undefined if p is either false or undefined at w. The idea has been revived and extended to be fit with modals (von Fintel, 2007; Huitink, 2009). It is problematic, though, since it seems to be saddled with the prediction that indicatives presuppose (entail, actually) their antecedents. That seems a distance from what conditional information is all about. (See also the discussion of this in von Fintel, 2012a, §5.2.)

References

Adams, E. W. 1975. *The Logic of Conditionals*. Dordrecht, Netherlands: Reidel.

Anderson, A. R. 1951. "A note on subjunctive and counterfactual conditionals." *Analysis*, 12(2): 35–38. DOI:10.1093/analys/12.2.35.

Austin, J. L. 1956. "Ifs and cans." *Proceedings of the British Academy*, 42: 109–132. DOI:10.2307/2964530.

Belnap, N. D. 1970. "Conditional assertion and restricted quantification." *Noûs*, 4(1): 1–12.

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bradley, R. 2000. "A preservation condition for conditionals." *Analysis*, 60(3): 219–222. DOI:10.1093/analys/60.3.219.

Edgington, D. 1995. "On conditionals." *Mind*, 104(414): 235–329. DOI:10.1093/mind/104.414.235.

Edgington, D. 2008. "Conditionals." In *Stanford Encyclopedia of Philosophy*, edited by E. Zalta. http://plato.stanford.edu/archives/win2008/entries/conditionals/ (accessed September 7, 2016).

Edgington, D. 2014. "Indicative conditionals." In *Stanford Encyclopedia of Philosophy*, edited by E. Zalta. http://plato.stanford.edu/archives/win2014/entries/conditionals/ (accessed August 20, 2016).

Fine, K. 1975. "Critical notice of *Counterfactuals* by David Lewis." *Mind*, 84(335): 451–458.

von Fintel, K. 1998. "The presupposition of subjunctive conditionals." In *MIT Working Papers in Linguistics: The Interpretive Tract*, vol. 25, edited by U. Sauerland and O. Percus, pp. 29–44. Cambridge, MA: MIT Press.

von Fintel, K. 2001. "Counterfactuals in a dynamic context." In *Ken Hale: A Life in Language*, edited by M. Kenstowicz, pp. 123–152. Cambridge, MA: MIT Press.

von Fintel, K. 2007. "If: the biggest little word." http://mit.edu/fintel/gurt-slides.pdf (accessed August 20, 2016). Slides from a plenary address given at the Georgetown University Roundtable.

von Fintel, K. 2012a. "Conditionals." In *Semantics: An International Handbook of Natural Language Meaning*, edited by K. von Heusinger, C. Maienborn, and P. Portner, pp. 1515–1538. Berlin: Walter de Gruyter.

von Fintel, K. 2012b. "Subjunctive conditionals." In *Routledge Companion to the Philosophy of Language*, edited by G. Russell and D. Graff Fara, pp. 466–477. New York: Routledge.

von Fintel, K., and A. S. Gillies. 2010. "Must…stay…strong!" *Natural Language Semantics*, 18(4): 351–383. DOI:10.1007/s11050-010-9058-2.

von Fintel, K., and A. S. Gillies. 2014. "Hedging your ifs and vice versa." Cambridge, MA: MIT/New Brunswick: Rutgers University.

von Fintel, K., and A. S. Gillies. 2015. ">= □ (+/− a bit but mostly +)." Cambridge, MA: MIT/New Brunswick: Rutgers University.

Gärdenfors, P. 1982. "Imaging and conditionalization." *The Journal of Philosophy*, 79(12): 747–760.

Gärdenfors, P. 1986. "Belief revisions and the Ramsey test for conditionals." *The Philosophical Review*, 95(1): 81–93.

Gibbard, A. 1981. "Two recent theories of conditionals." In *Ifs*, edited by W. Harper, R. Stalnaker, and G. Pearce, pp. 211–248. Dordrecht, Netherlands: Reidel.

Gillies, A. S. 2004. "Epistemic conditionals and conditional epistemics." *Noûs*, 38(4): 585–616. DOI:10.1111/j.0029-4624.2004.00485.x.

Gillies, A. S. 2007. "Counterfactual scorekeeping." *Linguistics and Philosophy*, 30(3): 329–360. DOI:10.1007/s10988-007-9018-6.

Gillies, A. S. 2009. "On truth-conditions for *if* (but not quite only *if*)." *The Philosophical Review*, 118(3): 325–348. DOI:10.1215/00318108-2009-002.

Gillies, A. S. 2010. "Iffiness." *Semantics and Pragmatics*, 3(4): 1–42. DOI:10.3765/sp.3.4.

Gillies, A. S. 2012. "Indicative conditionals." In *Routledge Companion to the Philosophy of Language*, edited by G. Russell and D. Graff Fara, pp. 449–465. New York: Routledge.

Grice, H. P. 1975. "Logic and conversation." In *Syntax and Semantics*, edited by P. Cole and J. L. Morgan, pp. 41–58. Cambridge, MA: Academic Press.

Groenendijk, J., and M. Stokhof. 1991. "Dynamic predicate logic." *Linguistics and Philosophy*, 14(1): 39–100.

Hájek, A. 1989. "Probabilities of conditionals – revisited." *Journal of Philosophical Logic*, 18(4): 423–428. DOI:10.1007/BF00262944.

Hájek, A. 1994. "Triviality on the cheap?" In *Probability and Conditionals: Belief Revision and Rational Decision*, edited by E. Eells and B. Skyrms, pp. 113–140. Cambridge: Cambridge University Press.

Harel, D., D. Kozen, and J. Tiuryn. 2000. *Dynamic Logic*. Cambridge, MA: MIT Press.

Herzberger, H. 1979. "Counterfactuals and consistency." *Journal of Philosophy*, 76(2): 83–88.

Huitink, J. 2009. "Domain restriction by conditional connectives." http://semanticsarchive.net/Archive/zg2MDM4M/Huitink-domainrestriction.pdf (accessed August 20, 2016).

Huitink, J. 2012. "McGee's counterexample to modus ponens in context." In *Studies in Meaning and Structure*, edited by B. Stolterfoht and S. Featherston, pp. 169–185. Berlin: Walter de Gruyter.

Jackson, F. 1991. *Conditionals*. New York: Oxford University Press.

Karttunen, L. 1973. "Presupposition and linguistic context." *Theoretical Linguistics*, 1(1–3): 181–194.

Kaufmann, M., and S. Kaufmann. 2015. "Conditionals and modality." In *Handbook of Contemporary Semantic Theory*, 2nd edn., edited by S. Lappin and C. Fox, 237–269. Hoboken, NJ: Wiley-Blackwell.

Khoo, J. 2013. "A note on Gibbard's proof." *Philosophical Studies*, 166(1): 153–164.

Kolodny, N., and J. MacFarlane. 2010. "Ifs and oughts." *The Journal of Philosophy*, 107(3): 115–143.

Kratzer, A. 1979. "Conditional necessity and possibility." In *Semantics from Different Points of View*, edited by R. Bäuerle, U. Egli, and A. von Stechow, pp. 117–147. Berlin: Springer.

Kratzer, A. 1981a. "The notional category of modality." In *Words, Worlds, and Contexts. New Approaches in Word Semantics*, edited by H.-J. Eikmeyer and H. Rieser, pp. 38–74. Berlin: Walter de Gruyter.

Kratzer, A. 1981b. "Partition and revision: the semantics of counterfactuals." *Journal of Philosophical Logic*, 10(2): 201–216.

Kratzer, A. 1986. "Conditionals." *Chicago Linguistics Society*, 22(2): 1–15.

Kratzer, A. 1989. "An investigation of the lumps of thought." *Linguistics and Philosophy*, 12(5): 607–653.

Kratzer, A. 1991. "Modality." In *Semantics: An International Handbook of Contemporary Research*, edited by A. von Stechow and D. Wunderlich, pp. 639–650. Berlin: Walter de Gruyter.

Kratzer, A. 2012. *Modals and Conditionals*. Oxford: Oxford University Press.

Ladusaw, W. 1980. "On the notion of 'affective' in the analysis of negative polarity items." *Journal of Linguistic Research*, 1: 1–16.

Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.

Lewis, D. 1975. "Adverbs of quantification." In *Formal Semantics of Natural Language*, edited by E. Keenan, pp. 3–15. Cambridge: Cambridge University Press.

Lewis, D. 1976. "Probabilities of conditionals and conditional probabilities." *The Philosophical Review*, 85(3): 297–315.

Lewis, D. 1979a. "Counterfactual dependence and time's arrow." *Noûs*, 13(4): 455–476.

Lewis, D. 1979b. "Scorekeeping in a language game." *Journal of Philosophical Logic*, 8(1): 339–359.

Lewis, D. 1980. "Index, context, and content." In *Philosophy and Grammar*, edited by S. Kanger and S. Öhman, pp. 79–100. Dordrecht, Netherlands: Reidel.

Lewis, D. 1981. "Ordering semantics and premise semantics for counterfactuals." *Journal of Philosophical Logic*, 10(2): 217–234.

Lewis, D. 1986. "Probabilities of conditionals and conditional probabilities II." *The Philosophical Review*, 95(4): 581–589.

McGee, V. 1985. "A counterexample to modus ponens." *Journal of Philosophy*, 82(9): 462–471.

Morreau, M. 2010. "It simply does not add up: trouble with overall similarity." *Journal of Philosophy*, 107(9): 469–490.

Moss, S. 2012. "On the pragmatics of counterfactuals." *Noûs*, 46(3): 561–586. DOI:10.1111/j.1468-0068.2010.00798.x.

Nute, D. 1980. *Topics in Conditional Logic*. Dordrecht, Netherlands: Reidel.

Pollock, J. L. 1976. *Subjunctive Reasoning*. Dordrecht, Netherlands: Reidel.

Ramsey, F. P. 1990 (1929). "General propositions and causality." In *Philosophical Papers: F. P. Ramsey*, edited by H. A. Mellor, pp. 34–51. Cambridge: Cambridge University Press.

Rothschild, D. 2012. "Do indicative conditionals express propositions?" *Noûs*, 47(1): 49–68. DOI:10.1111/j.1468-0068.2010.00825.x.

Segerberg, K. 1989. "A note on an impossibility theorem of Gärdenfors." *Noûs*, 23(3): 351–354.

Sobel, J. H. 1970. "Utilitarianisms: simple and general." *Inquiry*, 13(1–4): 394–449.

Stalnaker, R. 1968. "A theory of conditionals." In *Studies in Logical Theory*, edited by N. Rescher, pp. 98–112. Oxford: Blackwell.

Stalnaker, R. 1974. "Pragmatic presuppositions." In *Semantics and Philosophy*, edited by M. Munitz and P. Unger, pp. 197–213. New York: New York University Press.

Stalnaker, R. 1975. "Indicative conditionals." *Philosophia*, 5: 269–286.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, R. 2005. "Conditional propositions and conditional assertions." In *New Work on Modality (MIT Working Papers in Linguistics 52)*, edited by J. Gajewski, V. Hacquard, B. Nickel, and S. Yalcin. Department of Linguistics and Philosophy, MIT.

Starr, W. B. 2014. "A uniform theory of conditionals." *Journal of Philosophical Logic*, 43(6): 1019–1064. DOI:10.1007/s10992-013-9300-8.

Swanson, E. 2012. "Conditional excluded middle without the limit assumption." *Philosophy and Phenomenological Research*, 85(2): 301–321. DOI:10.1111/j.1933-1592.2011.00507.x.

Tichy, P. 1976. "A counterexample to the Stalnaker–Lewis analysis of counterfactuals." *Philosophical Studies*, 29(4): 271–273.

van Benthem, J. 1996. *Exploring Logical Dynamics*. Stanford: CSLI Press.

van Benthem, J. 2001. "Correspondence theory." In *Handbook of Philosophical Logic*, vol. 3, 2nd edn., edited by D. Gabbay and F. Guenthner, pp. 325–408. Dordrecht, Netherlands: Springer.

van der Does, J., W. Groeneveld, and F. Veltman. 1997. "An update on might." *Journal of Logic, Language, and Information*, 6(4): 361–380. DOI:10.1023/A:1008219821036.

van Fraassen, B. C. 1980. "Review of Brian Ellis, *Rational Belief Systems*." *Canadian Journal of Philosophy*, 10: 497–511.

Veltman, F. 1976. "Prejudices, presuppositions, and the theory of counterfactuals." In *Amsterdam Papers in Formal Grammar: Proceedings of the 1st Amsterdam Colloquium*, edited by J. Groenendijk and M. Stokhof, pp. 248–281. Amsterdam: Amsterdam University.

Veltman, F. 1985. "Logics for Conditionals." PhD thesis, University of Amsterdam.

Veltman, F. 1996. "Defaults in update semantics." *Journal of Philosophical Logic*, 25(3): 221–261. DOI:10.1007/BF00248150.

Veltman, F. 2005. "Making counterfactual assumptions." *Journal of Semantics*, 22(2): 159–180. DOI:10.1093/jos/ffh022.

Weatherson, B. 2009. "Conditionals and indexical relativism." *Synthese*, 166(2): 333–357. DOI:10.1007/s11229-007-9283-5.

Willer, M. 2014. "Dynamic thoughts on ifs and oughts." *Philosophers' Imprint*, 14(28): 1–30. http://hdl.handle.net/2027/spo.3521354.0014.028 (accessed August 20, 2016).

18

Generics

BERNHARD NICKEL

Generics (a linguistic phenomenon) exhibit genericity (not obviously a linguistic phenomenon), and though a theory of generics is closely connected to a theory of genericity, the two are distinct. They raise a host of interesting linguistic and philosophical issues, both separately and in their interaction.

1 Generics

Let's begin with a fairly manifest phenomenon we can observe in natural language. There is a range of sentences that, speaking intuitively, we can use to talk about kinds. The clearest examples are sentences that predicate properties that only kinds can have (what we may call *kind-restricted predicates*).

(1) Ravens are widespread.
(2) Dodos are extinct.
(3) Diamonds are rare.

Individual ravens can't be widespread; individual dodos can only be dead; and individual diamonds are all, by their very nature, unique.[1] Since (1)–(3) concern kinds – *genera* – the claims themselves have come to be called generics.

In addition to describing a kind using kind-restricted predicates, we can also characterize it, or perhaps characterize its members *qua* members of that kind, by using properties that individuals can have. We can call these latter generics *characterizing* generics. (4)–(12) are some true characterizing generics.

(4) Ravens are black.
(5) Tigers have stripes.
(6) Coke bottles have short necks.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

(7) Bishops move along diagonals.
(8) Lions have manes.
(9) Lions give birth to live young.
(10) Lions have four legs.
(11) Sea turtles are long-lived.
(12) Barns are red.[2]

We can contrast them with (13)–(17), which are false.

(13) Ravens are white.
(14) Prime numbers are odd.
(15) Supreme Court Justices have odd Social Security Numbers.
(16) Green bottles have short necks.
(17) Sea turtles die young.

As these examples indicate, a lot of our knowledge of the world is most naturally expressed using generic sentences. Generic sentences are also an important component of the language young children hear as they mature ("motherese").[3] For that reason, a semantic theory for natural language needs to have an account of generics.

These examples share several more specific features – beyond the intuitive sense that they are closely connected to kinds – that have made them of interest to philosophers and linguists alike, and these specific features support the intuitive sense that generics are particularly closely concerned with kinds.

1.1 Statistical Variability

There is no simple statistical criterion that systematically tracks the truth or falsity of characterizing generics. If we wanted to formulate a statistical criterion, the intuitive sense that generics characterize the kind at issue would suggest the hypothesis that they are universal generalizations. But empirically, that hypothesis is a non-starter, since most generics are compatible with a range of non-conforming cases. (4) and (5) are true in the face of albino ravens and tigers, respectively, while the truth of (8) tolerates the existence of female and immature male lions, and that of (11) the fact that a great many sea turtles are eaten by predators before they reach old age. A more plausible hypothesis holds that a generic is true if and only if most members of the kind have the property at issue. This hypothesis seems to be quite successful at capturing the intuitive force of (4) and (5), but it founders on examples like (11), since these are true even though most turtles die very shortly after birth, and on combinations like (8) and (9), since at most one of the males or females can be in the majority.

In response, we might further weaken the analysis of generics to require only that *some* members of the kind have the property at issue in order for the generic to be true, that is, analyze generics as existentials. This weak analysis is compatible with the truth of all of the examples (4)–(12), but it goes too far. It falsely predicts that *ravens are white* and *sea turtles are short-lived* are both true. For that reason, there is no single, simple statistical condition that successfully distinguishes true from false generics.[4] Even worse, a statistical approach just seems to be on the wrong track entirely when we consider generics that reflect rules such as (7). The truth of such a generic shouldn't be hostage to how often people happen to move bishops in some other way than along diagonals, be it in jest, out of carelessness, or for any number of intuitively irrelevant reasons.

1.2 Interlude: Generics and Context Dependence

In arguing that there's no simple statistical criterion that systematically captures the patterns of the truth and falsity of generics, I've so far ignored the possibility of invoking context-dependence. Yet it is well-known that many, and perhaps all, explicitly quantified statements are context sensitive. In different situations, they quantify over different domains. One might therefore wonder whether the problems generics raise dissipate as soon as we allow ourselves to draw on mechanisms of context sensitivity in interpreting generic sentences, which after all seem to express some sort of generality.

The problem does not dissipate so quickly. Even if generics are context sensitive, they are not context sensitive in the same way as explicitly quantified sentences. One salient difference concerns the features of the context that can influence the interpretation of the two sorts of sentences. Making a restriction extremely salient does not *ipso facto* influence the interpretation of generics as it does that of explicit quantifiers.[5]

Circus The lion act involves fierce lions that have to be restrained with chair and whip. The bear act involves bears riding unicycles and dancing.

(18) The bears are tame. (True)

(19) Bears are tame. (False)

So if we want to draw on mechanisms of context sensitivity as part of a theory of genericity, we must do so in a way that takes account of the distinctive features of genericity. I'll return to this issue throughout the chapter.

1.3 Modal Import

Generics seem to not just be substantially independent of statistical facts in the world and at the time with respect to which they are evaluated. They also seem to have implications for worlds or times beyond. We can see this by noting that some generics are false even though all members of the kind conform to the generalization, while others are true even though no members of the kind conform. A famous example of the first sort is (15): even if it is true that all Supreme Court Justices conform to (15), we still do not take it to be true. And the situation does not change if we know that no further Supreme Court Justices will be appointed, perhaps because the institution will be abolished when the next Justice steps down. There is no stronger statistical condition than that all members of a collection have the property predicated, so more than just the statistical facts that obtain at the world and time with respect to which (15) is evaluated must be relevant to its truth.

On the flipside, we find certain generics to be true even though no members of the kind conform to them. A standard example is (10). Not only is (10) true, it would still be true even if every single lion were to have some number of legs other than four, perhaps by some combination of mutation, accident, combat, and the machinations of a madman. But of course the truth of (10) in such a situation cannot be accounted for simply by suggesting that generics are equivalent to universal generalizations and that the universal generalization *all lions have four legs* is vacuously true. That would make *lions have seventeen legs* true as well. So more than just what is true at the world and time with respect to which (10) is evaluated must be relevant to its truth. These two examples illustrate the *modal import* of generics.

1.4 Well-Established Kinds

The sense that generics are peculiarly concerned with kinds is also sometimes supported by pointing to the contrast between (6) and (16), between *Coke bottles have short necks* and *green bottles have short necks*. While the former has a clear meaning and is indeed true, the latter is decidedly odd. This contrast is then explained by suggesting that generics can only concern well-established kinds, and that the green bottles fall short of constituting such a well-established kind.

I doubt, however, that this can be the right explanation for the contrast, since there are many generics that are clearly true but which do not concern kinds that can possibly count as more well-established than the green bottles. Consider (20).

(20) (Some true generics not about kinds)
   a. Albino ravens are white.
   b. Calico cats are female.

So some other account of the contrast is required, and it remains to be seen whether that contrast supports the sense that generics are closely concerned with kinds or whether it is simply an epiphenomenon.

In light of these observations, we should say that generics reflect an intimate connection between a property and a kind *or collection*, the connection that obtains when the property is characteristic of the kind or *collection*. But for ease of exposition, I'll continue to simply speak of generics as concerning kinds, taking it as read that they can also concern (at least some) non-kind collections.

Before moving on, let me emphasize that these theoretically interesting features are not meant as *tests* for genericity. Sentences that we have every reason to expect to be generics might not exhibit one or more of these features. An easy example is (21).

(21) Lions are mammals.

While the initial examples could at least potentially have counter-instances – members of the kind that do not have the property predicated in the generic – (21) could not. Presumably that is not due to some special feature of the semantic interpretation of generics, but rather to the fact that the characteristic connection happens to hold between a property and a kind that are also related by necessity.

2 Genericity

As can be seen from this list of theoretically important features of generics, the theoretically most central cases are characterizing generics, even though generics that predicate kind-restricted properties are the most obviously kind-directed forms of speech. For that reason, let's call *genericity* the phenomenon exhibited by characterizing generics: there is a particularly close or intimate connection between a kind and a property, one that does not obviously coincide with either statistical notions (all, most, many, some), nor does it coincide with well-established modal notions (necessity or essence). Let's call this connection *characteristic*.

Just about all theories of generics are, in the first instance, theories of genericity. They are a theory of the characteristic connection between a property and a kind that we see

expressed by characterizing generics. Since the meaning and use of generics is probably the most direct evidential connection we have to genericity, it's only natural to frame a theory of genericity as a theory of the truth-conditions of generics.

However, it's useful to separate generics from genericity, at least conceptually. We can see this by looking more closely at their relationship.

3 Separating the Semantics of Generics from Theories of Genericity

3.1 The Syntax–Semantics Interface

As a first point, we should note that there is no straightforward correlation between syntactic properties of a sentence and whether it can be used as a generic. All of the sentences that, in English, can express genericity, can also be used to express something non-generic.

The three forms of sentences that can be used as characterizing generics are bare plurals, singular definites, and singular indefinites. The basic paradigm is given in (22)–(24).

(22) Ravens are black.
(23) The raven is black.
(24) A raven is black.

Contrast these uses of the sentences as generics with their non-generic counterparts (25)–(27).

(25) Ravens are sitting on my lawn.
(26) The raven took the carrot.
(27) A raven scratched the post last night.

(25) and (27) are simply existential, (26) is about a particular bird.

There are some slight but striking differences among the different sentences we can use as generics. The singular definite suggests a stricter connection between property and kind, as the contrast between (28) and (29) indicates.

(28) Dutchmen are good sailors.
(29) The Dutchman is a good sailor.

(28) was first proposed in the *Port Royal Logic* of 1663.[6] At that time, it would certainly have been true to say it, even though (29) might have struck us as too strong.

(24) also has a stronger flavor than the bare plural (22). What is more, while both the bare plural and the singular definite happily combine with kind-restricted predicates, the singular indefinite does not, as (30) and (31) witness.

(30) *A raven is widespread.
(31) A raven is extinct.

(30) is simply impossible, while (31) is acceptable, but only has a reading that differs from its minimal counterpart *ravens are extinct*: the former can only mean that a subspecies of ravens has gone extinct, not that the whole species has. A semantic theory for generics needs to explain how these differences in the meaning of generics arises, but we probably won't learn much of interest about *genericity* from that.

The importance of separating generics and genericity comes out even more clearly when we look at one of the most prevalent arguments in discussions of genericity, arguments about the logical form of generics.

3.2 The LF of Generics

Let's call the kind of representation of a sentence that contains all of the information required for the interpretation of a sentence the LF of that sentence, where the expression 'LF' is supposed to be reminiscent of the notion of a logical form.[7] While sentences with more than one quantifier are ambiguous with respect to their scope, the LFs of such a sentence are unambiguous, to give just one example.

Two Options: Quantificational and Kind-Predicating Analyses

There are two camps that all accounts of the LF of generics fall into, and proponents of both camps model the LF of generics on that of other kinds of sentences. One camp looks to quantifiers as the crucial model, the other to referring expressions such as proper names. These hypotheses have a lot of empirical content because on either one, we immediately predict that generics are subject to the constraints that apply to the kinds of sentences they are modeled on, be that model quantificational or referential. These are predictions we can test, and they point to problems that proponents of either camp have to overcome in order to arrive at a more persuasive theory.

On a quantificational approach to generics, these sentences lack a quantifier only in so far as no quantifier is pronounced or written down. The LF of these sentences contains a quantificational element that has a distinctively generic meaning but is otherwise just like other quantificational expressions. The quantificational approach comes in two flavors, corresponding to two main ways that quantification is expressed in English. One models the generic quantificational element in a sentence such as (32a) on quantifiers such as *all* in (32b).

(32) a. Ravens are black.
b. All ravens are black.
c. Ravens are always black.

This determiner version of the generic operator is defended by some theorists.[8] An alternative version of the generic operator is defended by more, this one modeled on adverbs of quantification, such as *always*, *sometimes*, and *mostly*, illustrated in (32c).[9]

The second main approach to the LF of generics analogizes them to referring expressions. It is certainly the minority position.[10] It's motivated in the first place by drawing on the analogy between sentences in which a property is predicated of an object and sentences in which a kind-restricted property is predicated of a kind, an analogy exhibited in (33) and (34).

(33) Mary is home.
(34) Diamonds are rare.

Since being rare can only coherently be predicated of kinds, it seems extremely plausible that *diamonds* in (34) refers to a kind, just like *Mary* in (33) refers to a person. But then the null-hypothesis should be that *diamonds* uniformly refers to the kind, even in characterizing generics such as (35).

(35) Diamonds are valuable.

In addition to considerations of theoretical simplicity, proponents of the kind-predicating approach usually cite the problems the quantificational approach has as the main evidence in favor of their position.

Arguments that Turn on Scope

One famous argument turns on the scope possibilities generic sentences exhibit. Consider (36)–(38).

(36) Jane believes that some professors are eccentric.
  a. … just on general principle.
  b. … even though she doesn't know they're professors.
(37) Jane believes that Alice is eccentric.
(38) Jane believes that professors are eccentric.

When an ordinary quantified sentence is embedded in a propositional attitude verb such as *believes*, it can usually be interpreted in two ways. On the narrow scope interpretation, the ascription requires that the subject of the attitude, Jane in this case, has a general, quantified thought, and that she makes use of the concept picked out by the quantifier, *some professors* in this case. This reading is made prominent by the continuation (36a), which suggests that there aren't any particular professors Jane has in mind, she just thinks that there are some eccentric ones. On another reading, made prominent by the continuation (36b), Jane believes of some specific people that they are eccentric, and we as onlookers ascribing the belief to Jane make use of the notion of a professor, though Jane need not. This latter reading is apt when Jane meets what are in fact eccentric professors at an event without knowing that they are professors, but whom she quickly identifies as eccentric. The first reading is called the *narrow scope reading*, because it results from interpreting the quantified expression within the scope of *believes*, much as the sentence is actually sounded. The latter reading is the *wide scope reading*, since it corresponds roughly to *some professors are such that Jane believes them to be eccentric*.

There is no corresponding ambiguity in the interpretation of referring expressions that occur in the scope of propositional attitude verbs. (37) only has one reading, on which there is a specific person that Jane has in mind, though perhaps Jane thinks of the person in a different way. Generics seem to pattern with referring expressions, rather than quantified expressions, in this respect. (38) does not seem to be ambiguous. It can only mean that Jane believes the kind constituted by professors to be characterized by eccentricity. That suggests that (38) is associated with an LF on which *professors* refers to a kind, since we then predict the similarity in available and unavailable readings between (37) and (38).

The situation isn't completely univocal. Proponents of the quantificational analysis have offered the rejoinder that there *are* scope ambiguities with generics. Three important examples are (39)–(41).[11]

(39) Storks have a favorite nesting area.
  a. … which other storks try to capture.
  b. … which is in my backyard.
(40) John intentionally put belladonnas into the fruit salad because he mistook them for cherries.
(41) Typhoons arise in this part of the Pacific.

(39) has two readings, each of which we can bring to prominence with a suitable continuation. (39a) is apt if *a favorite nesting area* has narrow scope with respect to *storks*, (39b) if it has wide scope. (40) is intended to show that bare plurals can take wide scope with respect to propositional attitude ascriptions. Suppose that John is not trying to harm the people for whom he prepared the fruit salad, and keep in mind that belladonnas are poisonous. In that case, it cannot be part of his intention to put belladonnas in the salad. Rather, the intention was to put these things, which were in fact belladonnas but which John mistook for ordinary cherries, into the salad. And that proposition is expressed only if *belladonnas* takes wide scope with respect to the propositional attitude predicate *put intentionally*.[12] Finally, on a kind-referring strategy, we miss a possible reading of (41). On that strategy, the sentence is taken to mean that it is characteristic for typhoons to arise in this part of the Pacific. Typhoons arising elsewhere is an exceptional circumstance. But the much more prominent reading is that it's characteristic of this part of the Pacific that typhoons arise there, though typhoons arise in other parts of the world, as well.

This second set of examples suggests that the interpretive options for generics more closely mirror those of quantified sentences, not those of referential ones. The scope-taking characteristics of generics are thus equivocal, some pointing towards a kind-predicating analysis, some to a quantificational analysis. To resolve this issue, we need to look to further evidence. But we can already see the sorts of considerations that bear on the LF of generic sentences: they all concern whether positing one or another sort of LF for generics allows us to capture systematic constraints on interpretation.

This point becomes clearer still once we look at the most important challenges to a quantificational analysis of generics, which turn on how conjunctions work in generics.

Arguments that Turn on Conjunctions

Conjunctions raise several related problems for quantificational theories. The usual conclusion drawn from these problems is that only kind-predicating can meet the challenge. The purpose of this section is to articulate the challenge.

The first challenge, and one that is discussed quite frequently, centers on generics in which a property of kinds and a predicate of individual members of a kind are conjoined.[13] Consider (42) and (43).

(42) Diamonds are rare.
(43) Diamonds are valuable.

On a kind-predicating analysis, both sentences have structurally identical LFs, differing only in the predicate, given in (44) and (45), respectively.

(44) [diamonds rare]
(45) [diamonds valuable]

In both, *diamonds* denotes the kind, and both attribute a property to that kind. Using a semi-formal notation to mirror the LFs as much as possible, we can give the truth-conditions of (42) and (43) as (46) and (47), respectively.[14]

(46) $R(d)$
(47) $V_G(d)$

The predicate in (47) is subscripted to indicate that it is one that felicitously applies to kinds and is true just in case the kind is characterized by being valuable. Consider now the conjunction (48).

(48) Diamonds are rare and valuable.

The LF of this sentence is just like those of (42) and (43), except that the predicate now has more internal structure. Using the bracket notation, we can describe the LF as (49).

(49) [diamonds [rare and valuable]]

This LF is unproblematically interpretable, since the verb-phrase simply consists of a complex expression denoting a similarly complex property of kinds. We can state the predicted truth-conditions in a way that reflects this fact about the interpretability of (49) by making use of the notion of a conjoined predicate, as in (50).[15]

(50) $[R \wedge V_G](d)$

Here, $R \wedge V_G$ is a predicate true of a kind just in case that kind is both rare and characterized by being valuable.

Things appear differently for quantificational views. On a quantificational view, the initial examples (42) and (43) have very different LFs. The former predicates a property of the kind as a whole, so that we can simply retain the LF (44). But the LF of (43) is more complex.[16] For definiteness, assume that the generic operator *gen* occupies the same position as explicit quantifiers like *all* or *some*. The LF of (43) is thus (51).

(51) [[*gen* diamonds] [are valuable]]

To mimic the LF in a statement of the truth-conditions, we can give (52) as the output of the compositional system for (51).[17]

(52) [GEN x: $D(x)$]($V(x)$)

For all that I have said here, the truth-conditions (52), as predicted by the quantificational view, might be completely equivalent to the truth-conditions (47) as predicted by the kind-predicating view in the sense that they are true in all the same worlds. The crucial difference between the two views is how the truth-conditions are arrived at.

We should highlight a fact about quantificational LFs such as the one in (51). Like all generalized quantifiers, the generic operator has a restrictor and a scope, and the content of these two components is determined *structurally*, that is, the content of these two components is determined as a function of the syntactic structure of the LF of the sentence we're interpreting. The sister of the generic operator provides the restrictor, and the VP provides its scope.[18]

With this commitment in mind, consider how one can analyze the conjoined generic (48) on a quantificational approach. Obviously, there are two options: it can be given an LF that does not contain a generic operator, instantiating the same structure as (49), or it can be given an LF that *does* contain a generic operator, given in (53).

(53) [[*gen* diamonds] [rare and valuable]]

But on a quantificational approach to generics, neither LF gives rise to the proper interpretation. (49) is just uninterpretable, since according to it, a complex predicate is predicated of a kind-denoting expression, but that complex predicate does not pick out a property of kinds, because *valuable* does not pick out a property of kinds.[19]

The alternative LF (53) is no better, since it predicts that the sentence has the truth-conditions (54).

(54) $[\text{GEN}\, x\text{: } D(x)](R(x) \wedge V(x))$

That is to say, it requires that generically many diamonds are, each individually, both rare and valuable. And just as *valuable* cannot coherently be predicated of kinds, so *rare* cannot be predicated of non-kinds. Either way, (48) is predicted to be no good. But in point of fact, it is true. So quantificational approaches to generics must be rejected.

In giving this argument, I assumed for definiteness that the generic operator *gen* appears in the same position as quantificational expressions like *all* or *some*. Once this assumption is made, it follows from the basic structure of the sentence that the verb-phrase expresses a single complex condition, either by expressing a complex predicate that is predicated of the kind, as in (49), or by expressing a complex condition that is mapped to the scope of the generic quantifier and that individual members of the kind have to satisfy, as in (53). Opting for the alternative that the generic operator occupies not the position of a quantificational determiner such as *all* or *some*, but of an adverb of quantification such as *usually* or *normally*, does not change the argument substantially. For on this way of implementing the quantificational approach, it is still true that the whole VP expresses a single, complex condition that individual members of the kind need to meet. This argument against quantificational views of generics only relies on the basic architecture of the sentence. The argument requires no assumptions about how to interpret a generic operator. Therein lies its great strength.

But one might also think that combinations like *rare and valuable* are themselves quite rare and thus perhaps should not force quantificational theorists to abandon their position very quickly. Exceptional cases can perhaps be treated on an *ad hoc* basis. For this reason, I want to now look at another problem conjunctions raise, one which relies on slightly stronger assumptions about quantificational approaches to generics. These assumptions show the problem to be absolutely ubiquitous.

Conjunctions with More Complex Scope

This argument requires a bit more of a windup. Let me give you an intuitive statement first. An ordinary characterizing generic, such as *ravens are black*, divides the members of the kind into at least two categories (and if it's false, three). There are the members of the kind that the generic is *about* in some very rough sense, in this case, the black ravens that, so to speak, come by their blackness in the ordinary way (not ravens that are born as albinos and then dyed). There are the members of the kind that the generic is *not about*, such as the albinos. Those are the members that we feel are irrelevant when we're evaluating the truth of a generic.

What's more, different generics about one and the same kind divide the kind up in different ways. Consider (55).

(55) a. Ravens have two wings.
b. Ravens have two legs.

Still speaking intuitively, ravens that have lost a leg by accident, predation, or what have you, support the truth of (55a) just as much as ravens with two legs, so long as they have two wings. Ravens that have lost a wing by accident, predation, or what have you, support the truth of (55b) just as much as ravens with two wings, so long as they have two legs. These two generics, then, carve up the kind in different ways. But now consider (56).

(56) Ravens have two legs and two wings.

On a quantificational approach, (56) has an LF that predicts the following truth-conditions for it: generically many ravens have the complex property of having two legs and two wings. And in that case, the conjunction (56) is about fewer ravens than either of its conjuncts. In the case of (56), this may not be too bad. But consider (57).

(57) a. Lions have manes.
b. Lions give birth to live young.
c. Lions have manes and give birth to live young.
d. Lions have manes and lions give birth to live young.

(57a) excludes, *inter alia*, all of the female lions. (57b) excludes, *inter alia*, all of the male lions. On the quantificational approach, it would appear, (57c) therefore excludes all of the male *and* all of the female lions, that is, all of the lions. Yet there is clearly a reading of (57c) on which it says no more or less than (57d), the simple conjunction of (57a) and (57b).

Let me articulate this dialectic somewhat more precisely. In doing so, we'll have a chance to observe a fact about generics that is important in its own right: that they are inferentially inert in various ways. In fact, we can begin right there.

Even if it's true that, necessarily, all *Fs* are *Gs*, it still doesn't follow from the fact that *As are F* that *As are G*. A famous example of this inferential failure is the following.

Chicken Failure

(i) Chickens lay eggs.
(ii) Necessarily, all chickens that lay eggs are hens.
∴ (iii) Chickens are hens.

Yet on a quantificational approach, we'd expect that the inference is valid so long as the generic quantifier is upward entailing in its nuclear scope, an assumption all implementations of a quantificational approach accept. For illustration, assume that the generic quantifier means "all normal." In that case, premise (i) would mean that all normal chickens lay eggs; by premise (ii), that would entail that all normal chickens are hens. And that is just the interpretation of the conclusion (iii).

The standard response to this problem concerning inference is to suggest that the restriction of the generic quantifier depends on the predicate at issue, and perhaps on the context, so that different restrictions are at issue in the premise (i) and conclusion (iii). Just to give an example, if we thought that generics were about normal members of a kind, then on this revision, different notions or perhaps respects of normality would be at issue in premise and conclusion.[20]

The inferential profile of generics requires us therefore to posit different restrictors for the generic operator, depending on the predicate in the nuclear scope. With this in mind, return to an example I mentioned before, repeated here.

(58) Lions have manes and give birth to live young.

If we posit a generic quantifier that occupied either the position of a determiner, analogous to *all*, or an adverb of quantification, analogous to *always*, we would predict that all of these sentences only have interpretations that are clearly inappropriate, either because they are

false, or because they don't coincide with our intuitive sense of their meaning. The reason for this prediction lies in the very fact that already raised the trouble in interpreting (48): the whole predicate is interpreted as expressing a single, conjoined condition the kind is supposed to be characterized by. Hence, (58) is interpreted as expressing the proposition that generically many lions both have manes and give birth to live young, or put in terms of normality, that all normal lions have manes and give birth to live young. But such lions are extremely *ab*normal. This is the only reading a quantificational account seems to be able to assign to (58), and it therefore misses its most natural reading, one that is well paraphrased by *lions have manes and lions give birth to live young*.

I find these arguments the most powerful ones about the LF of generic sentences, far more than the inconclusive ones relating to scope I reviewed before. The data here are completely univocal.[21]

Before moving on, I want to point out a likely form of context dependence in generics. We have already seen that on a quantificational approach to generics, the restrictor of the generic quantifier must be determined at least in part by the predicate at issue. That was the lesson of the CHICKEN FAILURE. But that is probably insufficient to account for some contrasts, such as the one exhibited in (59).

(59) a. Bears ride on unicycles.
b. Bears ride in train cars.

Some semantic theories of generics predict that the predicates in (59a) and (59b) determine the same restriction of the generic quantifier, since both are concerned with locomotion. Such theories might appeal to context to determine a more fine-grained restriction, perhaps distinguishing between locomotion that is part of entertainment and locomotion that is part of transportation.

3.3 Separating the LF of Generics from Theories of Genericity

I have spent a great deal of time arguing about whether quantificational approaches to generics are correct. I now want to argue that the issue of whether a quantificational or kind-predicating approach is correct is orthogonal to the central issues raised by the phenomenon of genericity. Among these is perhaps the most basic: whether it's possible at all to give an informative account of genericity in the first place.

It's important to be clear on this because if there was a direct connection, we could investigate metaphysical questions with distinctively semantic means, since the arguments concerning the linguistic analysis of generic sentences invoke facts that are metaphysically neutral, such as scope possibilities and the interpretive options for conjunctions. The line of argument I have in mind begins with an analogy with explicit quantificational expressions. Consider a simple, universally quantified sentence, such as (60).

(60) All ravens are black.

The truth of that sentence supervenes in an obvious way on facts about individual ravens. (60) is true just in case every raven, perhaps in some contextually restricted range, is black (I'll ignore the contextual restriction from now on). The sentence is obviously quantificational, containing an element that denotes universal quantification, to wit, *all*. And when we give the compositional semantics for (60), we also give an informative analysis of that quantificational element, for example in terms of Tarski's semantics

for quantification. Indeed, the analysis of (60) that is part of the compositional semantics for English *just is* a description of the relationship between a collection and a property distinctive of universally quantified sentences: all members of the collection must have the property at issue.

Likewise, if generics are analyzed quantificationally, they contain an element that denotes generic quantification: *gen*. But then we also need to give it a semantics, and these semantics need to tell us how the truth of a generic about a kind depends on facts about individual members of that kind. If that's right, then at least the following holds: if we are skeptical of the possibility of giving an informative theory of genericity, then we're *ipso facto* skeptical of the possibility of stating the semantics of *gen*. But without a semantics, *gen* cannot make a systematic contribution to the meaning of sentences in which it appears, and hence generics don't have proper meanings on a quantificational approach to their interpretation. Hence, such skepticism should move us to accept that the interpretation of generics is kind-referring, not quantificational.[22]

However, drawing this sort of inference is not at all obligatory. Positing an element in the LF of a sentence is one thing. Being able to offer an informative analysis of that element is another. Indeed, drawing such a distinction is inescapable in the context of semantics as I've presented it here, since the point of this sort of semantics is to describe a relation between elements with a meaning that is taken as primitively understood – words, to a good first approximation – and elements with a meaning that is determined by them – paradigmatically sentences.

The point is perhaps most easily appreciated by considering simple predicates, such as *being a spy*. If I want to represent the semantically relevant information of the sentence *Lucy is a spy*, I will, *inter alia*, make use of a predicate *spy*. As far as the interpretation of that sentence is concerned, that is all that the occurrence of *spy* contributes. In the semantic interpretation, *spy* is then assigned an extension, perhaps the set of spies. As a separate matter, we may also believe that being a spy isn't the sort of property that floats freely in the metaphysics of the world. Rather, there's an informative theory of what makes someone a spy to be given. But whether or not there is such a non-trivial theory of spyhood is simply irrelevant to the question whether our LF representation is accurate or not. As far as the systematic description of the meaning of sentences containing that predicate is concerned, we can and perhaps should take it as primitively understood.[23]

The upshot of the discussion so far is that metaphysical issues about genericity and semantic issues about generics are independent in one direction. Denying the possibility of an informative theory of genericity is compatible with adopting either a quantificational or a kind-predicating LF for generics.

The independence holds in the other direction, as well. Accepting the possibility of giving such a theory of genericity is also compatible with adopting either sort of LF. The argument is straightforward. Suppose that we have a theory of genericity. This can be part of a semantic theory for a generic quantifier, precisely parallel to the semantics of the universal quantifier. But if it turns out that from the perspective of available readings and so on, generics behave as if they made reference to kinds, then the LF of generics should be kind-referring, and the theory of genericity is simply a non-semantic analysis of what characterizing a kind amounts to in exactly the same way that a theory of spyhood is a non-semantic analysis of what being a spy amounts to. Hence, there is no straightforward substantive connection between debates about the LF of generics and the possibility of describing genericity generally and informatively.

4 Connecting the Semantics of Generics with Theories of Genericity

I have argued that there is no direct connection between a linguistic semantic theory of generics and a broadly metaphysical theory of genericity. But that is not to say that there is no connection at all. There are many important ones.

In the first instance, our use of generic sentences is our best and most direct evidence for what genericity is like, and constructing a semantic theory of generics will help us to distinguish good from bad evidence in the usual way in so far as we'll be able to say more explicitly and precisely how much of our use of generics reflects the semantic content of generics and how much is a pragmatic accretion. In this respect, generics are exactly like all other phenomena where our intuitions about the world, as expressed in language, provide us with evidence for our theorizing.

But in the second instance, a semantic theory for generics plays a more distinctive role, as well. Genericity is, as I've been at pains to emphasize, a theoretical notion. We introduce it by reflecting on various phenomena, such as statistical variability and modal import, and positing a common, unifying explanation for these phenomena. But then we can also ask which phenomena, exactly, need to be accounted for by a theory of genericity. Once again, let me consider a particular example in order to make the issue vivid, (28), repeated here.

(28) Dutchmen are good sailors.

How can we fit the interpretation of (28) into a more general account of generics?

Set aside for the nonce the choice of formal tools and approach we want to bring to bear on our analysis of generics. Let's focus only on the intuition a formal treatment could be designed to capture. The most popular and prominent examples of characterizing generics, those that initially shape a lot of theories, are (61)–(63).

(62) Ravens are black.
(62) Tigers have stripes.
(63) Lions have manes.

These examples suggest that the characteristic connection between property and kind is a somehow close or intimate one.

But this motivating intuition seems to be undermined by the fact that there are some generics where it seems incredibly implausible that there could be a close connection between belonging to a kind and the property predicated in the generic. (28) is just such an example. Even at a time when the Dutch had a maritime empire it didn't seem as if there was anything that made the very good sailors particularly good examples of being Dutch. (64) is another example of the same type.

(64) Ravens are bigger than toasters.

It just seems odd to think that there's anything about being a raven that puts ravens in a close relationship to properties of toasters. They certainly aren't selected for being bigger than toasters (no toasters under conditions of selection) nor are toasters part of raven ecology today.

There thus seems to be a real tension, simply at the level of motivating intuitions. On the one hand, the examples (61)–(63) suggest that genericity is a close or intimate connection,

an intuition further supported by the modal import we've already noted. On the other hand, (28) and (64) suggest that the connection can be quite loose, and certainly not essential-ish in the way suggested by the previous examples. How should we respond to this sort of tension?

We could deny that (28) needs to fit into a theory based on the guiding intuition that characteristic connections are close connections. Perhaps generics are simply ambiguous, with one interpretation being exemplified by *ravens are black*, another by *Dutchmen are good sailors*. Such a denial could take the form of an appeal to context sensitivity, in this case, contextual variation in the sort of relationship that is said to hold between a property and a kind.

Consider, for example, generics from the contested and controversial realm of politics. Based on the high statistical correlation between being a kindergarten teacher and being a woman, a speaker who holds strong biological determinist views may assert (65) to express these views.

(65) Kindergarten teachers are women.

However, another speaker might also point out that there are various social factors, including upbringing, discrimination, and so on, that bring it about that many women are steered towards less prestigious and less well-paid professions. At the same time, certain professions, including kindergarten and preschool teacher, are perceived as "women's work," and hence paid relatively little. So it is far from an accident that the statistical fact – that the vast majority of kindergarten and preschool teachers are women – obtains. Against this background, an opponent of the sorts of injustice I just mentioned may also wish to assert (65) to mark out precisely that there is a very strong and important connection between being a kindergarten teacher and being a woman, albeit one that is external to the individual agents involved, and one that we should oppose.[24] Two speakers who disagree about the underlying social, political, and biological facts may thus convey very different propositions with the same generic. The question for such an invocation of context-dependence is whether it can be sufficiently constrained so as to avoid making it too easy for generics to be true.[25]

A second way to respond to the initially puzzling cases (28) and (64) suggests that the initial intuition that generics are always about a very close, almost essential, connection between property and kind, was mistaken. In some cases in which the characteristic connection holds, an intimate or quasi-essential connection holds as well, as is perhaps the case in the motivating examples, but that does not make the characteristic connection the quasi-essential one. Here, too, the threat of over-generation lurks.

A third option is easier to see for (64). It may turn out that the apparently problematic fact is nothing more than an interaction effect between genericity and other, semantically significants features of the generic sentences involved.

In the case of (64), the problem arises because the sentence appears to predicate *being bigger than toasters* of the kind raven and hence claims that this property is the characteristic one. But that may be mistaken. An alternative linguistic analysis of the sentence holds this: ravens have a certain characteristic size, toasters have a certain characteristic size, and the former size is greater than the latter. In that case, what's characteristic of ravens is not a relation to toasters, but simply having a certain size. And there may well be a close or intimate connection between ravens and their characteristic size.

So the construction of a semantic theory of generics is almost certainly an important part of coming to a satisfactory account of genericity, but how the two are related must be determined on a case-by-case basis.

5 Ascriptions of Dispositions, Habits, and Capacities

I have so far suggested that we should, at least conceptually, distinguish between generics and genericity, and that genericity can manifest itself when we assert or think that there is a characteristic connection between a property and a kind. A natural question is whether genericity is exhausted by this characteristic connection.

The matter is unclear. Consider another linguistic phenomenon that appears to share many of the theoretically important features we noted in re: characterizing generics. These are ascriptions of dispositions, habits, and capacities, examples of which are given in (66).

(66) a. Jane drinks coffee. (habit)
b. My Peugeot goes 120 mph. (capacity)
c. This glass breaks when struck. (disposition)

There are many similarities between these ascriptions and the generics I've been discussing so far that justify treating them all as belonging to a single class. In an ascription of a habit, disposition, or capacity – a *habitual* for short – we take a property that is most directly true of something at a specific time, such as drinking right here and now, being driven at 120 mph yesterday morning from eight to nine, or breaking when the glass was struck last week, and we ascribe that property to objects in a less direct way, so that the predicate can be true of the object even while the object isn't actively engaged in, say, drinking, running, or breaking. We thus have a parallel to the intuitive sense of "moving up" in category that we see when we take a property true in the first instance of objects and ascribe it to a kind made up of that sort of object.

Just like characterizing generics, ascriptions of habits, capacities, and dispositions exhibit an independence from the particular situations with respect to which they are evaluated. They can be true, even if the object to which they are ascribed never exhibits the property ascribed, as in the famous (67).

(67) Mary sorts the mail from Antarctica.

It may be true that sorting Antarctic mail is Mary's job, in which case (67) is true, even if no mail ever arrives from Antarctica and Mary perforce never sorts any such mail.

In light of these similarities, it is certainly plausible to think that habituals and characterizing generics are both manifestations of the same underlying phenomenon, that is, that genericity can manifest itself both in our thought and talk about characteristic connections, and in our thought and talk about dispositions and so on.

Regardless of whether we want to give a uniform account of characterizing generics and ascriptions of dispositions, habits, and capacities, another question arises. What is the relationship between ascriptions of dispositions and habits, which seem to exhibit some sort of genericity, and ascriptions of dispositions that do not make use of generics, such as *the glass is fragile*? It may be possible, for example, to analyze all ascriptions of dispositions in terms of generics.[26] The answer to this question is currently a matter of debate.

6 Some Theories of Genericity

To give a sense of the theoretical options available, I will briefly review four approaches to generics and genericity.

6.1 Is There a Logic of Generics?

In order to frame the discussion, I have treated genericity as a more or less metaphysical notion – genericity is something that manifests itself in the characteristic connection between a property and a kind. But in the first instance, one may think that genericity is an aspect of our thought and talk, so that it is a feature of the objects of our beliefs and assertions. It's therefore an open possibility that genericity should be understood less as a feature of the world, and more as a feature of how we use language in making our way around the world. One such approach analyzes generics in terms of their logic.

We have already seen that generics don't enter into deductive relations with other generics in connection with the failed inference concerning chickens. A generic sentence doesn't validly entail anything about individual members of a kind. By way of illustration, we may observe that the following inference is invalid, since Tweety might be a penguin.

Tweety Inference

(i) Bids fly.
(ii) Tweety is a bird.
∴ (iii) Tweety flies.

Further, we cannot validly conclude anything about supersets or subsets from the truth of a given generic, as illustrated by (68) and (69).

(68) Baby giraffes have short necks.
⇏ Giraffes have short necks.
(69) Lions have manes.
⇏ Female lions have manes.

The former shows that we cannot generally draw an inference from a subset to a superset, the latter that we cannot generally draw an inference from a superset to a subset. There therefore aren't any distinctively generic, fully general deductively valid inferential patterns.

Now, while the Tweety inference isn't valid, it isn't completely terrible either. In very many circumstances, the conclusion receives a lot of support from the premises, although a speaker or thinker may be forced to withdraw the conclusion if she acquires further information, for example, that Tweety is a penguin. Nonetheless, it does seem that it's acceptable to provisionally draw the Tweety inference until such extra information becomes available.

A formal treatment of such provisional inferences is given by non-monotonic logics: suppose we have an inference relation $\vdash$ that relates sets of sentences and sentences. The logic characterized by $\vdash$ is *non-monotonic* just in case the following generalization fails for at least some set of sentences Γ and sentences φ, ψ: if $\Gamma \vdash \varphi$ then $\Gamma \cup \psi \vdash \varphi$. Informally: adding premises to a good inference does not make it bad.

One may hold that certain non-monotonic inferences exhaust the meaning of generic sentences. This is the non-monotonic analogy to the idea that certain deductive inferences are constitutive of the meaning of the logical constants. The meaning of conjunction is exhausted by pointing to the inferences it underwrites, *and*-introduction and *and*-elimination. Likewise, the meaning of the generic *ravens are black* is completely captured by the fact that a speaker who accepts it is licensed to infer that something is black if she learns that it's a raven, and she has no information to the contrary that defeats the inference.[27]

A central question for such an approach concerns just what counts as additional information that may defeat the inference. The problem arises most clearly for generics that are true, even though only a minority of members of the kind conform to it. Consider (70).

(70) Sea turtles are long-lived.

Assume, as is indeed the case, that most sea turtles die young. Should we treat this knowledge as defeating the default inference that Sandy is long-lived, knowing only that Sandy is a sea turtle? On the one hand, the usual gloss of knowledge that defeats an inference is knowledge that is specific to the case at hand, such as that Tweety is a penguin. On the other hand, the information that most sea turtles die young is additional information that goes beyond the generic sentence itself.[28]

6.2 Probabilistic/Majority-Based Approaches

An approach to theorizing about genericity that has a similar starting point but divorces it from actual inferences focuses on probabilities. It is due to the work of Ariel Cohen (1999a; 1999b; 2003; 2004). Cohen's account is an interesting example of a theory that is designed to account for various important features of generics by positing an interaction between a compositional semantic theory for generics and a non-semantic, metaphysical theory of genericity, couched in terms of probability.

Most abstractly, he suggests that a characterizing proposition is true just in case *most* members of the relevant kind *in a suitable domain* have the property at issue. For Cohen, all of the work thus takes the form of specifying the suitable domain in a general and informative way.

We can see these two components – majorities and the suitability of the domain – in action by seeing how Cohen's semantics account for some basic data.

(71) a. Ravens are black.
b. Ravens are white.

On Cohen's view, (71a) is true just in case the probability that a randomly chosen raven is black is greater than even. Within Cohen's system, that is equivalent to the claim that most ravens in a suitable reference class are black. Assuming for now that all of the ravens that exist form a suitable reference class, Cohen's view correctly predicts that (71a) is true. Likewise, (71b) requires that most ravens in that class are white. Since that is not the case, Cohen's view correctly predicts that (71b) is false.

Of course, if we simply consider the distribution of a property among members of a kind at the world and time of evaluation, generics are predicted to be susceptible to what are intuitively irrelevant accidents. Cohen handles these sorts of blips by taking a suitably long view of the matter. Though at a moment the actual frequency with which a property is instantiated in a population might vary widely, over time that frequency will tend towards a more stable value. It is the stable, long-term value that is relevant to the evaluation of a generic, not the potentially fluctuating short-term value.

We now have the raw materials of Cohen's view, but not quite the whole semantics, since the materials I've introduced so far aren't quite sufficient to account for examples such as (72).

(72) a. Lions have manes.
b. Lions give birth to live young.

Cohen diagnoses the reason that the examples in (72) are true as the result of restricting our attention to a proper subset of the members of a kind. He accomplishes that restriction by introducing alternatives: whenever a generic predicates a property of a kind, we have to interpret that generic with a range of implicitly given alternatives to the property actually predicated. In the case of (72a), these alternatives are properties such as having colorful tail-feathers, having large antlers, and other forms of sexually selected ornamentation; in the case of (72b), the alternatives include giving birth by laying eggs, by cell division, and so on. When we evaluate a generic, we only look at those members of a kind that satisfy at least one of the alternatives. In (72a), that has the effect of excluding all of the females and immature males from consideration, since they don't have any forms of sexually selected ornamentation; in (72b), it has the effect of excluding males and infertile females, since they don't reproduce by any means. Among the remaining members of the kind, it is indeed the case that most members conform to the generalization, so that (72a) and (72b) are correctly predicted to be true.

To summarize, we can state Cohen's truth-conditions for generic sentences.[29]

(73) Let [GEN x: $A(x)$]($F(x)$) be a sentence, where A and F are properties. Let ALT(F) be the set of alternatives to F. Then
[GEN x: $A(x)$]($F(x)$) is true iff $P(F|A \wedge \vee \text{ALT}(F)) > 0.5$

In words: *As are F* is true just in case, in a suitable set of *A*s, all of which have at least one of the alternatives to *F*, most are *F*.

These truth-conditions show the interaction between non-generic semantic facts, specifically, the provision of a set of alternatives that are modeled on the interpretation of focus, and a theory of genericity, which is reflected in the majority-based quantification and in the specification of a suitable domain.

6.3 *Cognitive Approaches*

A very different idea is embodied in cognitive approaches which take generics to quite directly reflect aspects of our cognitive system. These include theories according to which generics reflect stereotypes (e.g., Declerck, 1986; Geurts, 1985) and approaches on which they reflect conceptual connections (e.g., Prasada and Dillingham, 2006; 2009). One central question for such approaches concerns how to handle generics that don't concern lexicalized concepts, that is, generics that concern collections that do not form a kind and that do not plausibly have associated stereotypes, such as (74) and (75).

(74) Cats with blue eyes are usually blind.
(75) Whoopee cushions left on stuffy professors' chairs on Friday afternoons are funny.

It also seems as if such cognitive approaches can be too divorced from the world. Consider a baseless stereotype. How can we ensure that the corresponding generic comes out false, rather than true?

An interesting development of a cognitive view that avoids some of these problems has been presented by Sarah-Jane Leslie (2008). On her view, some generics track relatively objective facts about the world, others much more subjective ones. *Ravens are black*, for example, tracks the fact that we take there to be characteristic sorts of properties that organisms have, including coloration, and black is the characteristic color for ravens. Other generics simply reflect that certain facts are deeply impressive to us as subjects in the world,

including shark attacks and the transmission of viruses by mosquitos, which is why *sharks attack bathers* and *mosquitos carry the West Nile Virus* are true. The central question for an account like this is whether the psychological facts really do coincide with the truth-conditions of generics: is it the case, for instance, that every fact in the world that is striking, dangerous, or otherwise impressive is also honored with a corresponding generic?

Cognitive approaches to generics generally, and Leslie's work in particular, raise a host of foundational questions about the relationship between semantics and human cognition. Leslie argues for her theory, for example, on the grounds that it fits more naturally with observations about language acquisition, a source of evidence semantic theories usually do not consider. Her work also suggests that the sorts of patterns that are usually considered as best accounted for with the tools of formal semantics are perhaps better handled by appeal to more idiosyncratic aspects of human psychology.

6.4 Normality-Based Approaches

A fourth family of views takes its inspiration from the intuition that generics are about what is normal. The intuition can be brought out most clearly, perhaps, by considering how we respond to members of a kind that do not conform to a given generic we accept as true. Confronted with an albino raven, we might say that, at least as far as *ravens are black* is concerned, the albino is abnormal. That response suggests that generics are about what is normally the case.[30] The key questions for such an account are twofold: What is the target of normality, and how can the notion of normality be spelled out substantively.

The first question concerns the choice whether we should interpret generics in terms of normal worlds, or in terms of a more fine-grained notion, such as normal individuals. As we saw in section §1, generics have modal import, so the theoretically most natural treatment of generics would be to interpret them in terms of normal worlds, perhaps using different notions of normality in different contexts. An initial implementation of such an approach would interpret a generic *As are F* as true iff in all most normal worlds, all *A*s are *F*. Unfortunately, this approach faces significant problems. To give a flavor of the concern, return to *sea turtles are long-lived*. A world in which all sea turtles are long-lived is a world in which sea turtles aren't subject to predation. But that is, speaking intuitively, not a biologically normal world.

A more natural approach may look towards a more fine-grained notion of normality, perhaps looking at normal members of a kind. This raises the second key question for normality-based approaches: What does it mean to be normal, given that it is not a statistical notion? What makes this question particularly vexing is that generics cover a very broad range of phenomena, including the natural world, the social world, rules of games (*bishops move along diagonals*), to name just a few. The question is whether it's possible to describe a notion of normality that is rich enough to explain why, for example, albino ravens aren't normal when we evaluate *ravens are black*, while being flexible enough to apply across all domains.

7 Two Ways of Doing Away with Genericity

I have been speaking of theories of genericity, assuming that genericity is a phenomenon in its own right. This assumption can be questioned, and I want to look at two different ways of doing away with it.

7.1 Falsity and Pragmatics

One might think of generics as very widespread instances of *loose talk*, a notion that has been described by Lasersohn (1999). And if generics really are instances of loose talk, the variability we see in their interpretation does not reflect a theoretically unified phenomenon of genericity – it reflects a hodgepodge of incredibly localized, pragmatic factors that aren't susceptible to systematic theorizing. To assess the merits of this concern, let me begin by introducing the notion of loose talk.

Consider an assertion of (76).

(76) Mary arrived at 3:00 p.m.

In very many circumstances, we're happy to accept the sentence – to not critique the speaker on the grounds that she has said something false – even if Mary didn't arrive at 3:00 p.m. on the dot, but only at a few minutes before or after 3:00. One way of theoretically describing what's going on holds that, though the sentence is false in such a situation, it conveys a truth because we were speaking somewhat loosely. The reason to think that the sentence is actually false but only conveys something true is that, when conjoined with the tolerated slack in the situations we'd accept as being described by a use of that sentence, we get a contradiction. *Vide* (77).

(77) Mary arrived at 3:00 p.m., though she arrived at 3:05 p.m.

More generally, instances of loose talk are cases in which a sentence is used that is false when evaluated with respect to a given circumstance, but where the context is such that the aspects of the situation that account for the sentence's falsity can be acceptably ignored.

An extension of these ideas into a treatment of generics may appear attractive. Consider, for example, the exchange (78).

(78) A: Birds fly.
B: What about penguins?
A: # Birds fly.

A's simple repetition in this exchange is odd, and this oddity is easily accounted for if generics were simply instances of loose talk. On this strategy, generics have the same truth-conditions as the corresponding universal generalizations, but the bare plural indicates that we are speaking somewhat loosely. *A*'s initial assertion is thus false – it is false that all birds fly – but it might nonetheless be unobjectionable if the flightless birds can be ignored. *B*'s question, and its concomitant introduction of penguins into the conversation, makes it so that they cannot be ignored any longer. Hence, the falsity of *A*'s second utterance can no longer be overlooked.

Appeals to loose talk also promise straightforward accounts of the very vexing examples that make arriving at a substantial theory of genericity so difficult. After all, it might make sense to ignore a vast number of members of a kind if they're all uniformly irrelevant. Hence, *sea turtles are long-lived* is handled easily. The sentence is false, but in most situations in which we care about the lifespan of sea turtles, we're ignoring what happens to them immediately after hatching. We only care about the ones that have made it to the pet store.

If (uses of) generics really are instances of loose talk, then there is no such thing as genericity. There is no single, underlying phenomenon that accounts for the properties of interest of generics across a wide range of cases. Rather, what we can ignore in a context

given our other aims and interests is just about guaranteed to be a sundry and miscellaneous mess, since the aims and interests we bring to different contexts are sundry and miscellaneous, precisely not the manifestation of a single, theoretically coherent phenomenon.

Happily for the prospects of an informative semantics for generics, the loose talk conception is not really tenable. For one, we've seen that a key reason to identify a use of a particular sentence as an instance of loose talk comes from the felt contradictoriness of conjoining that sentence with an explicit acknowledgment of the kind of looseness we're willing to tolerate. To make the point maximally vivid, it it useful to switch to another of Lasersohn's examples. He points out that we are often willing to accept (79) even though not quite all of the townspeople are asleep.

(79) The townspeople are asleep.

We might, for example, plan to raid the town, and we're happy to accept the sentence so long as we can legitimately ignore those townspeople who are still awake – perhaps the awake ones won't foil our plans. Now consider the contrast between (80) and (81).

(80) # The townspeople are asleep, though some townspeople are awake.
(81) Ravens are black, though some ravens are white.

The former is far worse than the latter. In fact, (81) is perfectly acceptable. That strongly suggests that generic sentences are true. We can see the same point by contrasting how loose talk interacts with negation. In an instance of loose talk, the sentence used is false, so that negating it yields a truth. Here too, loose talk and uses of generic sentences differ markedly.

(82) Mary didn't arrive at 3:00 p.m.
  a. …, she arrived at 3:05 p.m.
(83) Ravens aren't black.
  a. …, some ravens are white.

If Mary really arrived at 3:05, then (82) is perfectly acceptable, though perhaps pedantic. But (83) is clearly bad, even given the analogous continuation in (83a). Furthermore, the loose talk conception of generics can make no sense of the fact that some generics strike us as false even though the corresponding universal generalization is both true and acceptable, such as *Supreme Court Justices have odd Social Security Numbers*. For on the loose talk conception, that just means that the generic is true, as well. Our use of generics isn't simply an instance of loose talk.

7.2 Thoroughgoing Context Sensitivity

We have seen throughout this chapter that generics are context sensitive in several distinctive ways. Their domain appears to be restricted, albeit through different mechanisms than ordinary quantified sentences. One such mechanism may involve a restriction of the domain as a function of the predicate, as shown by the CHICKEN FAILURE and related phenomena. We have also seen that, at least in some cases, it may be plausible that which connection is said to hold by a generic can be sensitive to context, as may happen when discussing politically and socially contested generics.

One can take the existence of these forms of context-dependence as an indication that generics are thoroughgoingly context sensitive. Indeed, one could even claim that all of the facts that we initially took as manifestations of a single, underlying phenomenon, genericity,

are better explained as many different manifestations of context-dependence. This might be the case if, as Sterken (2015) has argued, the very force of the quantification involved in generics is determined by context. On such a view, it's not just that context determines a way for us to restrict our attention, as in *all contextually salient As are F*. Instead, context determines what sort of quantification we're countenancing, as in *Q many As are F*. On such a view, distinctive context-dependence is all there is to genericity. The key question is whether such an approach offers a more systematic account of generics than one that leans more heavily on a mix of semantics and metaphysics.

8 Closing

Writing in 2016, we can say that the study of generics is still in its very early stages. Fundamental questions are left open: Which phenomena should be treated together? Which separately? Which framework or frameworks are most promising?

Philosophers and linguists are also just beginning to investigate the possibilities of drawing on genericity as an explanatory notion in other arenas. In philosophy, that includes debates about dispositions, about the nature and status of *ceteris paribus* laws in the special sciences, the nature and resistance to countervailing evidence of stereotypes, and the normative domain, to name just a few.

Notes

1 One may well want to avoid complicating one's ontology by positing kinds in addition to other abstract objects such as properties. To do so, one could interpret predicates such as *widespread*, *extinct*, and *rare* as properties of the properties of being a raven, a dodo, and a diamond, respectively. However, this strategy leaves unexplained why the following sentence is infelicitous, not just false: *the students in this class are extinct*. Why can *extinct* be predicated of the property of being a dodo but not of the property of being a student in this class? The same point also shows that we cannot simply analyze such kind predication as collective predication. Predicates such as *be from different countries* are similar to kind-predicates in that they cannot be predicated of individuals. However, as the example I just gave shows, kind-predicates are much more selective than collective predicates, which suggests that they apply to a special range of objects: kinds.

2 This may strike some readers as an odd example, since redness is not anywhere near as intimately connected to barns as blackness is to ravens. But it is a widely attested generic, and examples like this are important data points for theories of generics and genericity. They challenge theories that attempt to account for the meaning of generics too directly in terms of an intimate or close connection between kind and property.

3 See, e.g., Gelman (2003).

4 I qualify this conclusion to only apply to "simple" statistical conditions since we'll see a more complex statistical account in discussing Ariel Cohen's (§6.2). His account is not so easily refuted.

5 This sort of case is reported in Krifka *et al.* (1995, p. 45). As with everything connected to generics, the facts are not univocal, as the following contrast illustrates.

(A) # Rooms are square.
(B) (Traditional Japanese architecture has strict rules.) Rooms are square.

While (A) is odd, in part because it is a claim about rooms generally, the context set up by (B) manages to restrict the interpretation of *rooms* to rooms that follow the rules of traditional Japanese architecture (cf. Greenberg, 2003, p. 36).

6 Arnauld and Nicole (1996/1683).
7 Nonetheless, I don't call it a logical form, because LFs in this context are a term of art that may well diverge from the thing or things philosophers call logical forms in other contexts.
8 See, e.g., Asher and Morreau (1995).
9 Lewis (1973) is the basic text in the contemporary study of adverbs of quantification. Cohen (1999a; 1999b), Schubert and Pelletier (1989), and Wilkinson (1991) are some of the theorists who treat generics on this model.
10 This view is most famously associated with Greg Carlson's dissertation (Carlson, 1977) and has recently been defended by Liebesman (2011).
11 (39) is from Schubert and Pelletier (1987), (40) from Kratzer (1980), and (41) from Milsark (1974).
12 A potentially confusing fact about the argumentation surrounding (40) is that *belladonnas* must be given an existential reading, rather than a generic, on a par with *ravens are sitting on the telephone wires*, which means that there are some ravens sitting on the wires, not that it's characteristic of ravens to sit there. But this difference between generic and existential interpretations is irrelevant for the debate about kind-referring and quantificational LFs for generics, since on the kind-referring strategy, *all* uses of bare plurals are kind-referring, be they existential or generic. Thus, showing that existential bare plurals do not behave as predicted on the kind-referring strategy simultaneously undermines the kind-predicating strategy for generic uses of bare plurals.
13 These examples were already urged in favor of a kind-predicating analysis in Carlson's dissertation (1977), and they have since been endorsed by Schubert and Pelletier (1987) and, very recently, Liebesman (2011).
14 Notice, however, that what's crucial *isn't* some particular way of stating the truth-conditions of these sentences. The issue turns on *how* the truth-conditions are assigned, specifically, on the LF that serves as the input to the compositional semantic machinery. Thus, the point of this specification of the truth-conditions is just to give a sense of the LFs involved.
15 To ease exposition, I couch the discussion of the compositional semantics in as informal terms as possible. In particular, I eschew λ-notation as much as possible. In that notation, the predicate would be $\lambda x\,.\,\left[R(x) \wedge V\,G(x)\right]$.
16 It's a further issue how the two kinds of LF are related. Perhaps generics are simply ambiguous, and one of the readings is ruled out on pragmatic grounds. Alternatively, one of the LFs might be basic and the second is generated in response to uninterpretability. For the second view, see Cohen (2001).
17 I'll use GEN, an expression in the metalanguage, as the interpretation of *gen*, a silent operator that is part of the object language.
18 Once we take into account further phenomena, such as focus, mapping the material in a sentence into restrictor and scope becomes harder, but for now, we can retain the simple hypothesis.
19 What exactly goes wrong may differ from one particular implementation to another. Perhaps the VP is uninterpretable because of a type-mismatch. Perhaps the VP is somehow interpretable, but then the sentence as a whole is ill-formed because of a mismatch between part of the meaning of the VP and the subject. The details don't matter to the argument.
20 In a quantificational setting, Asher and Morreau (1995), Cohen (1999a; 1999b), and Nickel (2008) have proposed similar theories. Gerstner-Link (1988) and ter Meulen (1986) offer a similar strategy in the context of interpreting generics as modalized conditionals. Finally, the idea that restrictions imposed by the predicated property are theoretically important is also present in the circumscription approach due to McCarthy (1980; 1986).
21 For some recent discussion, see Leslie (2015), and Nickel (2016).
22 This line of thinking is quite explicit in Carlson's dissertation (Carlson, 1977) which set the framework for most further theorizing about generics (Carlson, 1977, p. 73).
23 One theorist who I think is best interpreted in these terms is Sarah-Jane Leslie, especially her (2008). This is the contrast Leslie has in mind when she distinguishes the truth-conditions of a generic (a semantic notion) from worldly truth-makers (a metaphysical notion).

24 Cf. Haslanger (2012).
25 I will return to this point in §7.2, when I discuss ways of eliminating genericity.
26 Cf. Fara (2005).
27 In a dynamic setting, Veltman (1996) is a good discussion. In a completely different framework, Brandom (1994) is an instance of this strategy.
28 For some recent work on the connection between generics and default inference, see Pelletier (2010) and Pelletier and Elio (2005).
29 Adapted from Cohen (1999b, p. 37).
30 See, for example, Asher and Morreau (1995), Eckardt (1999), Nickel (2008; 2016) and Schurz (2001).

References

Arnauld, A., and P. Nicole. 1996 (1683). *Logic or the Art of Thinking*. Cambridge: Cambridge University Press.

Asher, N., and M. Morreau. 1995. "What some generic sentences mean." In *The Generic Book*, edited by G. N. Carlson and F. J. Pelletier, pp. 300–339. Chicago: University of Chicago Press.

Brandom, R. B. 1994. *Making it Explicit*. Cambridge, MA: Harvard University Press.

Carlson, G. N. 1977. "Reference to Kinds in English." PhD thesis, University of Massachusetts.

Cohen, A. 1999a. "Generics, frequency adverbs, and probability." *Linguistics and Philosophy*, 22(3): 221–253.

Cohen, A. 1999b. *Think Generic!* Stanford, CA: CSLI Publications.

Cohen, A. 2001. "On the generic use of indefinite singulars." *Journal of Semantics*, 18(3): 183–209.

Cohen, A. 2003. "Existential generics." *Linguistics and Philosophy*, 27(2): 137–168.

Cohen, A. 2004. "Generics and mental representation." *Linguistics and Philosophy*, 27(5): 529–556.

Declerck, R. 1986. "The manifold interpretations of generic sentences." *Lingua*, 68(2–3): 149–188.

Eckardt, R. 1999. "Normal objects, normal worlds, and the meaning of generics." *Journal of Semantics*, 16(3): 237–278.

Fara, M. 2005. "Dispositions and habituals." *Noûs*, 39(1): 43–82.

Gelman, S. A. 2003. *The Essential Child*. Oxford: Oxford University Press.

Gerstner-Link, C. 1988. "Über Generizität. Generische Nominalphrasen in singulären und generischen Aussagen." PhD thesis, University of Munich.

Geurts, B. 1985. "Generics." *Journal of Semantics*, 4(3): 247–255.

Greenberg, Y. 2003. *Manifestations of Genericity*. New York: Routledge.

Haslanger, S. 2012. "Ideology, generics, and common ground." In *Resisting Reality*, pp. 446–477. Oxford: Oxford University Press.

Kratzer, A. 1980. "Die analyse des blossen plurals bei gregory carlson." In *Linguistische Berichte*, 70: 47–50.

Krifka, M., F. J. Pelletier, G. N. Carlson *et al.* 1995. "Genericity: an introduction." In *The Generic Book*, edited by G. N. Carlson and F. J. Pelletier, pp. 1–124. Chicago: University of Chicago Press.

Lasersohn, P. 1999. "Pragmatic halos." *Language*, 75(3): 522–551.

Leslie, S.-J. 2008. "Generics: cognition and acquisition." *Philosophical Review*, 117(1): 1–47.

Leslie, S.-J. 2015. "Generics oversimplified." *Noûs*, 49(1): 28–54.

Lewis, D. K. 1973. "Adverbs of quantification." In *Formal Semantics of Natural Language*, edited by E. L. Keenan, pp. 3–15. Cambridge: Cambridge University Press.

Liebesman, D. 2011. "Simple generics." *Noûs*, 45(3): 409–442.

McCarthy, J. 1980. "Circumscription: a form of non-monotonic reasoning." *Artificial Intelligence*, 13(1–2): 27–39.

McCarthy, J. 1986. "Applications of circumscription to formalizing common sense knowledge." *Artificial Intelligence*, 28(1): 89–116.

Milsark, G. 1974. "Existential Sentences in English." PhD thesis, MIT.

Nickel, B. 2008. "Generics and the ways of normality." *Linguistics and Philosophy*, 31(6): 629–648.
Nickel, B. 2016. *Between Logic and the World*. Oxford: Oxford University Press.
Pelletier, F. J. 2010. "Are all generics created equal?" In *Kinds, Things, and Stuff: Mass Terms and Generics*, edited by F. J. Pelletier, pp. 60–79. New York: Oxford University Press.
Pelletier, F. J., and R. Elio. 2005. "The case for psychologism in default and inheritance reasoning." *Synthese*, 146(1–2): 7–35.
Prasada, S., and E. M. Dillingham. 2006. "Principled and statistical connections in common sense conception." *Cognition*, 99(1): 73–112.
Prasada, S., and E. M. Dillingham. 2009. "Representation of principled connections: a window onto the formal aspect of common sense conception." *Cognitive Science*, 33(3): 401–448.
Schubert, L. K., and F. J. Pelletier. 1987. "Problems in the representation of the logical form of generics, plurals, and mass nouns." In *New Directions in Semantics*, edited by E. LePore, pp. 385–451. London: Academic Press.
Schubert, L. K., and F. J. Pelletier. 1989. "Generically speaking, or, using discourse representation theory to interpret generics." In *Properties, Types, and Meaning*, vol. II, edited by G. Chierchia, B. H. Partee, and R. Turner, pp. 193–268. Dordrecht, Netherlands: Kluwer Academic.
Schurz, G. 2001. "What is 'normal'? An evolution-theoretic foundation for normic laws and their relation to statistical normality." *Philosophy of Science*, 68(4): 476–497.
Sterken, R. K. 2015. "Generics in context." *Philosophers' Imprint*, 15(21): 1–30.
ter Meulen, A. 1986. "Generic information, conditional contexts, and constraints." In *On Conditionals*, edited by E. Traugott, A. ter Meulen, J. Snitzer-Reilly, and C. Ferguson, pp. 123–145. Cambridge: Cambridge University Press.
Veltman, F. 1996. "Defaults in update semantics." *Journal of Philosophical Logic*, 25(3): 221–261.
Wilkinson, K. 1991. "Studies in the Semantics of Generic Noun Phrases." PhD thesis, University of Massachusetts.

19

Deflationist Theories of Truth, Meaning, and Content

STEPHEN SCHIFFER

I

When a philosopher proposes a semantic theory she commends for being deflationist,[1] that theory is intended to replace an entrenched theory that is predicated in part on what is thought to be the need to answer certain questions, and the philosopher objects to that theory not because it gives the wrong answers to those questions, but because she feels those questions are the wrong ones to be asked by a theory seeking to explain what the entrenched theory might legitimately hope to explain. Her alternative theory, she contends, is both to be preferred to the entrenched theory and deflationist relative to it because it doesn't bear the burden of needing to answer those questions.

Every deflationist semantic theory has its inflationist correlate: this is the semantic theory the deflationist theory is designed to deflate. Consequently, distinct deflationist semantic theories may differ from one another either by having different inflationist correlates or by having the same inflationist correlate but differing in their views of how its deflating should go. Some deflationist semantic theories are considerably more deflationist than others. For example, a deflationist theory of truth as applied to propositions needn't be at all deflationist about semantic notions in their application to sentences or utterances. The most interesting deflationist theories would be ones that are deflationist about all semantic notions in all their applications; but of such theories I am aware of only one. I'll call this theory *Radical Deflationism*, and I'll take its inflationist correlate to be a theory I'll call *Radical Inflationism*, although it will be obvious that there are a number of inflationist alternatives to Radical Deflationism that are less inflationist than Radical Inflationism. I shall present Radical Inflationism and Radical Deflationism as stipulatively defined theories, without regard to who might subscribe to them, or to one or another of their parts, but, as will become clear as I proceed, Radical Deflationism is based on a view worked out over a number of important publications by Hartry Field.[2] This chapter is mostly structured around a discussion of Radical Deflationism, but Part II briefly discusses another way of being semantically deflationist.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

A. Radical Inflationism

The radical inflationist doesn't doubt that there are semantic and propositional-attitude facts, and she holds that semantic and propositional-attitude predicates such as 'means,' 'is true,' 'refers to,' 'is true of,' 'believes that,' and so on both apply contingently to the things to which they apply and are univocally applicable both interlinguistically and interpersonally. For example, she holds that the properties ascribed by the italicized predicates in

- 'Brutus killed Caesar' (as used by English speakers) *means that Brutus killed Caesar*
- 'Brutus killed Caesar' (as used by English speakers) *is true*
- 'Brutus' (as used by English speakers) *refers to Brutus*
- 'Killed' (as used by English speakers) *is true of Brutus and Caesar (in that order)*
- I *believe that Brutus killed Caesar*

are exactly the same as the properties they ascribe in

- 'Брут убил Цезаря' (as used by Russian speakers) *means that Brutus killed Caesar*
- 'Брут убил Цезаря' (as used by Russian speakers) *is true*
- 'Убил' (as used by Russian speakers) *is true of Brutus and Caesar (in that order)*
- Vladimir *believes that Brutus killed Caesar*.

She also holds, first, that semantic and propositional-attitude facts play indispensable causal-explanatory roles in explaining how we are able to use the utterances and beliefs of others as a source of information about the world, and in predicting and explaining human behavior, and second, that, because they play those causal-explanatory roles, our semantic and propositional-attitude notions must be explicable in terms of physically realizable functional or causal notions. In other words, the radical inflationist is on board with the view Hartry Field had when he wrote "Tarski's theory of truth," that our notions of reference and truth are correspondence notions that stand in need of physicalistic explications (Field, 1972).

It's not difficult to see the path that might lead a radical inflationist to the other extreme, the view I'm soon to call Radical Deflationism. First, she is apt to find that it proves "extraordinarily difficult to develop the details of an adequate correspondence theory" (Field, 1986, p. 67) – that is to say, extremely difficult to find plausible naturalistic explications of our semantic and propositional-attitude notions. Second, as we are about to see, Radical Deflationism avoids the need to find physicalistically acceptable underpinnings of any kind for our semantic and propositional-attitude notions, and the frustrated radical inflationist is apt to find very strong appeal in "the prospect of avoiding the need to work out the details of [a physicalistically acceptable explication of our semantic and propositional-attitude notions], by declaring that [the interpersonally applicable semantic and propositional-attitude notions she thought were in need of physicalistic explication] serve no very important purpose" (Field, 1986, p. 67). And third, while the appeal of that prospect wouldn't carry much weight for a radical inflationist unless she had reason to doubt whether the intentional notions in question served any very important purpose, she is apt to give credibility to Radical Deflationism's claim that it's "extremely difficult to find a persuasive argument" (Field, 1986, p. 67) that we *need* the semantic and propositional-attitude notions that define Radical Inflationism.

B. Radical Deflationism

In presenting Radical Deflationism it's advisable to proceed in stages. This is because the theory takes as its core semantic notions certain egocentric use-independent disquotational notions, but those notions aren't themselves able to do all that the radical deflationist thinks semantic notions are needed to do, so the theorist finds it necessary to define certain non-egocentric semantic notions in terms of those core notions. Consequently, I propose to conceptualize Radical Deflationism as developing in six stages. Each stage will have the radical deflationist *stipulatively define* the semantic and propositional-attitude notions he takes himself to need at that stage, until by the end of the sixth stage he has all the intentional notions needed to do what, by his lights, semantic and propositional-attitude notions are needed to do. At that point he will leave open the question of to what extent his stipulatively defined notions might serve to explicate the intentional notions we actually use, for, while he doesn't rule out that his notions might do a passable job of capturing what is legitimate in their ordinary-language counterparts, he has little-to-no interest in accounting for the use of 'means,' 'true,' 'believes,' and so on in ordinary language, a question he regards as being "of only sociological interest" (Field, 1994a, p. 133). His primary interest in the notions that are to define Radical Deflationism is to determine whether they can do the work we might legitimately expect intentional notions to do, thereby showing we have no *need* for inflationist semantic or propositional-attitude notions.

Stage One: Unambiguous Eternal Sentences

Let's pretend for a while that every sentence of *my* language is an unambiguous eternal sentence[3] that, unlike 'This sentence isn't true,' doesn't give rise to any liar-like antinomy. The predicates to be introduced at this stage are 'means_{ss},' 'true_{ss},' 'false_{ss},' 'refers_{ss},' and 'true-of_{ss}.' I intend the stipulations by which I will introduce them to make these predicates (nearly enough) what Field (in the works cited) would call "purely disquotational." The 'ss' subscript indicates that the subscripted predicates apply to all and only the expressions I "understand," as I "understand" them, where this implies, for example, that "if on my understanding of 'Der Schnee ist weiss' it is equivalent to '$E = mc^2$,' then for me this sentence is [true_{ss}] iff $E = mc^2$" (Field, 1994a, p. 134). I have 'understand' in scare quotes to indicate that the word is not being used as it's used in ordinary language, and certainly not as the radical inflationist would use it. On that, the inflationist use, for a person to understand an expression is for her to use, or to know how to use, it *correctly*, so that, for example, if I use 'Exercise is enervating' the way you use 'Exercise is energizing,' then I have used it incorrectly and thus *mis*understand it. But for the radical deflationist there is nothing to count as my misunderstanding an expression that has a use in my idiolect; if I use 'Exercise is enervating' the same way I use 'Exercise is energizing,' then that just shows that my understanding of the one sentence is the same as my understanding of the other. What the radical deflationist means by 'understanding a sentence' is roughly as follows. Assume that I think in a language of thought, which may be the same as the language I speak (whatever exactly that is taken to mean), and let's continue to pretend that the languages with which we are at this point concerned are ones all of whose sentences are unambiguous and eternal. Whether or not my mentalese is a neural version of my spoken language, every spoken sentence of mine will, relative to our simplifying assumptions, be uniquely correlated with a mentalese sentence whose conceptual role is the conceptual role my spoken sentence would have if my

spoken language were my language of thought, where a sentence's conceptual role is the role the sentence has in my perceptual belief formation and in my theoretical and practical reasoning. A sentence's conceptual role is determined by the conceptual roles of its syntactic structure and constituent morphemes, and those conceptual roles are the contributions those things make to the conceptual roles of the sentences in which they occur. A mentalese sentence may be prevented from playing its conceptual role if for example it's too long or convoluted for one to process owing to the computational limitations of one's brain. If we ignore those limitations, however, we can say, as for convenience I will say, that, for the radical deflationist, my understanding of a sentence just is the conceptual role of its mentalese correlate. If, consequently, the mentalese correlates of my sentences 'Exercise is enervating' and 'Exercise is energizing' have the same conceptual role, then no invidious comparison can be made between my understanding of the two sentences; they merely happen to be the same.

The stipulations that define Radical Deflationism at this first stage are as follows.

'*Means*$_{ss}$' applies only to expressions I understand, and every instance of the schema

'*S*' means$_{ss}$ that *S*

(e.g., '"Snow is white" means$_{ss}$ that snow is white') is for me *a priori* and empirically indefeasible.

'*True*$_{ss}$' and '*false*$_{ss}$' apply only to sentences I understand, and every instance of the schemas

'*S*' is true$_{ss}$ iff *S*
'*S*' is false$_{ss}$ iff not-*S*

is for me *a priori* and empirically indefeasible in a way that makes instances of

'*S*' is true$_{ss}$
'*S*' is false$_{ss}$

cognitively equivalent for me to their corresponding instances of

S
Not-*S*,

where to say that two sentences are "cognitively equivalent" for a person is to say that "the person's inferential procedures license a fairly direct [and 'more or less indefeasible'] inference from any sentence containing an occurrence of one to the corresponding sentence with an occurrence of the other substituted for it; with the stipulation … that the occurrence to be substituted for is not within the context of quotation marks or an intentional attitude construction" (Field, 1994a, p. 107, fn. 2).[4]

'*Refers*$_{ss}$' applies only to names I understand, and every instance of the schema

If *n* exists, then '*n*' refers to *n*

is for me *a priori* and empirically indefeasible. (Thus, my acceptance of the sentence '"Ned Block" refers$_{ss}$ to Ned Block' is entirely independent of whether or not anyone ever used 'Ned Block' to refer to Ned Block and of whether or not his parents gave him that name.)

'*True-of*$_{ss}$' applies only to predicates I understand, and every instance of the schema

'*F*' is true-of$_{ss}$ *o* iff *o* is *F*

is for me *a priori* and empirically indefeasible. Similarly for predicates of arity greater than one, although representing this in a schema would require special conventions for correlating instances of the right-hand side of

'R^n' is true-of$_{ss}$ $< o_1, \ldots, o_n >$ iff $R^n(o_1, \ldots, o_n)$

with sentences of my language, such as 'Jane gave her tiara to Oxfam.'

I trust it's clear that the foregoing stipulations have no implications at all as regards the truth-conditions of any sentence of my language, or as regards what any expression in my language means, refers to, or is true of, as 'means,' 'true,' 'refers,' and 'true of' are used by an English-speaking neutral observer of the inflationism–deflationism fray. Radical Deflationism gives every speaker of every language her own egocentric disquotational semantic notions. But knowing how, say, a speaker of Inuit uses her egocentric disquotational predicates would not enable you to infer anything about what her expressions mean, refer to, or have as their truth-conditions. This is especially brought home by the following observation. If one thinks of a language as an abstract object that may or may not be used by anyone – say, as a pairing of sounds and meanings over an infinite domain – then there are infinitely many distinct languages that share exactly the same expressions as English. For example, whereas in English

'Snow is white' means that snow is white
'Grass is green' means that grass is green
'Coal is black' means that coal is black

in another language, English*,

'Snow is white' means that coal is black
'Grass is green' means that snow is white
'Coal is black' means that butter is fattening.

If the expressions of a person's idiolect are the same as those of English and English*, then the question arises as to whether she speaks English, English*, or another one of the infinitely many languages whose expressions are shared with those two languages. Now suppose that my use of the expressions in my idiolect is such that any informed neutral observer would confidently claim that my language is English*. Notwithstanding that, it would remain the case that in my spoken language

'Snow is white' means$_{ss}$ that snow is white
'Grass is green' means$_{ss}$ that grass is green
'Coal is black' means$_{ss}$ that coal is black.[5]

A central claim of Radical Deflationism will be that we have no need of any inflationist semantic notions, and that the only semantic notions we need are ones definable in terms

of the core egocentric disquotational notions (semantic$_{ss}$ notions, in my case). But what need do we have for the radical deflationist's egocentric disquotational notions, notions that apply only to expressions one understands? I'm not aware that the radical deflationist has at hand an answer for every one of the disquotational notions he recognizes, especially for his notion of meaning$_{xx}$ that applies only to sentences one understands, but he does have an answer as regards his egocentric disquotational notion of truth (in my case, the notion of truth$_{ss}$). It's an answer perhaps first made explicit by Quine:

> [T]he truth predicate is superfluous when applied to a given sentence; you could just utter the sentence. But it is needed for sentences that are not given. Thus we may want to say that everything someone said on some occasion was true, or that all consequences of true theories are true. (Quine, 1992, p. 80)

Radical deflationists such as Stephen Leeds and Hartry Field understand Quine to be saying that the *only* reason we need a truth predicate is that it functions as a device for expressing certain infinite conjunctions and disjunctions,[6] and their point is that the purely disquotational egocentric notion of truth is sufficient to satisfy that need. For example, in my idiolect the sentences

(1) Ava said something true$_{ss}$
(2) Everything Bob said is true$_{ss}$

are equivalent, respectively, to

(3) Ava said 'Snow is white' and snow is white, or Ava said '68 + 57 = 5' and 68 + 57 = 5, or …
(4) If Bob said 'Snow is white,' then snow is white, and if Bob said 'Fleas have souls,' then fleas have souls, and …

This, however, needs qualification in at least three respects. First, as Field (1994a, p. 120, fn. 17) recognizes (and no doubt Leeds as well), (1) isn't equivalent to (3) and (2) isn't equivalent to (4): nothing in (3) or (4) entails that the sentences referred to in them are all the sentences of my idiolect, so it's logically possible for (3) to be false and (1) true by virtue of the fact that the only true$_{ss}$ sentences Ava uttered are not among the sentences referred to in (3), and it's logically possible for (4) to be true and (2) false by virtue of the fact that none of the false$_{ss}$ sentences Bob uttered are among the sentences referred to in (4). Second, when Quine wrote that the truth predicate "is needed for sentences that are not given" he had in mind not only one's need to say things like 'I managed to say something true in my lecture today,' where I am talking about my own sentences, but also such things as 'Not everything the Pope says when he's speaking *ex cathedra* is true,' when I don't understand a word of Latin. But of course, as so far described, the radical deflationist doesn't have the resources to say that. Third, it may be that while (1) and (2), *which use the stipulatively introduced predicate 'true$_{ss}$,'* are correct, they aren't useful characterizations of Ava or Bob. For it may be that, while (1) and (2), which use the predicate 'true$_{ss}$,' are correct,

(5) Ava said something true
(6) Everything Bob said is true,

which use the ordinary language predicate 'true,' are incorrect. For it may be that, while Ava's and Bob's sentences are the same as mine, they mean – now using 'mean' as its used in

ordinary language – quite different things for Ava and Bob than they do for me. The radical deflationist needs to introduce notions that will enable him to claim that his notions are able to confer the benefits that our ordinary semantic notions may legitimately claim to provide. This is our segue to the next stage, but a final point is in order before we get to it.

This is that Radical Deflationism's claim *that the only reason we need a notion of truth is as a device of disquotation* is hostage to the theory's ability to meet what we will presently see is its greatest challenge – namely, to show that it has the wherewithal to account for the role of content in propositional-attitude explanations. For suppose I explain Olga's petitioning to have her marriage annulled by revealing that she has recently come to believe *that when her husband married her he was already married to someone else*. Since a belief that *P* is a belief that is *true iff P*, I have apparently explained Olga's petitioning for an annulment in terms of her being in a state with a certain *truth-condition* – to wit, in terms of her being in a state that is true iff her husband was married to someone else when he married Olga. It's incumbent on the radical deflationist to show that he can account for the role of truth-conditions in propositional-attitude explanation using only a deflationist-licensed notion of truth, and this is an issue we have yet to consider.

Stage Two: Foreign Sentences

Our egocentric disquotational notions can't apply to foreign sentences we don't understand, but if the radical deflationist is to show we don't need inflationist semantic notions to get the benefits we actually get from the semantic notions we actually use, he will want to show how he can use his core egocentric disquotational semantic notions to define non-egocentric non-disquotational semantic notions to apply to foreign sentences we don't understand, and this because we will certainly want to be able to use the utterances and propositional-attitude-states of foreign language speakers as a source of information about the world and to predict and explain their behavior. How is that to be done?[7] However the radical deflationist tries to show it can be done, he must take care not to end up in effect defining the very inflationist "correspondence" semantic notions the avoidance of which is the very point of the position he is trying to develop. At the same time, as Field points out, if he does find himself realizing that he does need inflationist notions to accommodate foreign language utterances (or, for that matter, anything else), then that would have the unexpected benefit of proving to him that what he thought couldn't be done could be done. That is why Field's official position isn't Radical Deflationism but is rather what he calls *methodological deflationism*: the methodological policy that "we should start out assuming [Radical Deflationism] as a working hypothesis; we should adhere to it unless and until we find ourselves reconstructing what amounts to the inflationist's relation '*S* has the truth-conditions *p*'" (Field, 1994a, p. 119). Of course, Field wouldn't be justified in accepting methodological deflationism unless he thought there was a pretty good *prima facie* case to be made for thinking that the only semantic and propositional-attitude notions we need are those permitted by Radical Deflationism. The feasibility of methodological deflationism turns on whether there is a pretty good *prima facie* case to be made.

Field mentions two approaches the deflationist might take to accommodating foreign language utterances we don't understand. The first, which he says is the least satisfactory of the two approaches, is a weakening of Radical Deflationism that he calls "extended-disquotationalism." This would be "to use a notion of interlinguistic synonymy: where '*S*' is a foreign sentence I don't understand, regard '*S* is true' as equivalent to '*S* is synonymous

with a sentence of ours that is true in the pure disquotational sense'" (Field, 1994a, p. 128), where 'synonymy' is a relation defined in terms of certain of the conceptual-role and indication-relation features of my own words.[8] The challenge for the deflationist disposed to take this option, Field says, would be to define a notion of interlinguistic synonymy without relying on a prior notion of truth-conditions, and he is very skeptical that that can be done. I would have thought that an even bigger problem for extended-disquotationalism is that it's hard to see how defining a notion of interlinguistic synonymy wouldn't be tantamount to defining an inflationist notion of truth-conditions. For suppose the radical deflationist were to devise a notion of interlinguistic synonymy. Such a notion would doubtless be defined using conceptual-role and indication-relation features of my sentences; but whatever features of my words were used to define a notion of interlinguistic synonymy, they would have to be physicalistically acceptable, and for every two sentences *S* and *S′* of my language whose truth_{ss} conditions differed, there would have to be distinct properties φ and φ′ such that (i) *S* has *φ* and *S′* has *φ′* and (ii) a foreign sentence was for me synonymous with *S* iff it had *φ* and synonymous with *S′* iff it had *φ′*. Suppose such a definition of synonymy at hand. Then for every sentence *S* of my language, the predicate '___ means that *S*' will be interlinguistically applicable and will ascribe the same use-dependent property to every sentence of which the predicate is true. Wouldn't that be one kind of inflationist conception of meaning and, *a fortiori* (since a sentence's meaning that *S* entails its being true iff *S*), one kind of inflationist conception of truth-conditions? Field raises this question for a view he calls 'quasi-disquotationalism' apparently without noticing that extended-disquotationalism entails quasi-disquotationalism (Field, 1994a, p. 131).

The second option Field says is available to the radical deflationist is the one he thinks that theorist will need to adopt, so presumably it's an option he thinks might give the radical deflationist all that he can legitimately expect to gain from a way of assigning truth-conditions to foreign sentences he doesn't understand. The option is based on the notion of a foreign sentence's being "*true relative to a correlation* of it to one of our sentences" (Field, 1994a, p. 128), where:

> For any foreign language *L* and any correlation *C* of *L* sentences with my sentences, an *L* sentence *S* is *true relative to C* iff there is a sentence *S′* of mine such that *C* correlates *S* with *S′* and *S′* is true_{ss}.

The deflationist will then explain that:

> There is no such thing as a sentence I don't understand being *per se* true or false. Ways of correlating sentences of a foreign language with my sentences are in effect ways of translating that language into mine. There is no question of a translation being right or wrong, of being the *correct* or *incorrect* translation of a foreign sentence into my language, only of a translation's being better or worse in a way that is "highly context-dependent (since the purposes for which they are better or worse vary from one context to the next)" (Field, 1994a, p. 128). If I'm unable to interpret a foreign sentence, then "it makes no sense [for me] to inquire whether it is true (except perhaps in counterfactual terms about how I would interpret it under definite conditions)."[9]

This is pretty radical stuff. I don't understand a word of Arabic, but I would have thought that infinitely many Arabic sentences are true or are false irrespective of any way of

correlating them with my sentences, and therefore irrespective of how I choose to interpret them or whether or not it's even possible for me to interpret them; and I would have thought that if a speaker of English translated 'La neige est blanche' as anything other than 'Snow is white,' then he simply mistranslated it: the French sentence means in French what the English sentence means in English, and not to know that is not to know what at least one of the two sentences means. If, however, the only notion of truth I am able to apply to foreign sentences I don't understand is the one Radical Deflationism makes available to me, then I must accept that foreign sentences I don't understand have no truth-conditions, and therefore no truth-values – except relative to this, that, or the other way of correlating those sentences with mine. The departures from common sense snowball. For example, I have for a long time taken myself to have a considerable amount of general knowledge about speakers of, say, Japanese, even though when I hear Japanese speech I can't even tell where one word ends and another begins. I know that many Japanese have bought iPhones, that many have been tourists in the United States, that many have applied and gone to universities, that many get married, that many who get married later get divorced, that some join the army, that some quit their jobs, that some are orthopedic surgeons who specialize in knee replacements, and so on and on and on. But according to Radical Deflationism, it's impossible for me to know any of those things, and this because the theory entails that there can be no fact of the matter as to whether a speaker of Japanese bought an iPhone, quit her job, filed for divorce, performed a knee replacement, or did any of the myriad other things I thought I knew they did. There can be no fact of the matter about such things because to do any of them requires acting with intentions with particular contents and Radical Disquotationalism entails that there can be no fact of the matter about the contents of the propositional-attitude-states of speakers of languages we don't understand. For if we were able to assign objective contents to their propositional-attitude-states, then we would surely also be able to assign objective contents to the sentences they use to express those states.

Does the fact that Radical Deflationism entails these departures from common sense constitute an *objection* to it? Not if it can be shown that all the real benefits of having our commonsense conception of truth could just as easily be had if each person used only her own egocentric disquotational notions of truth and falsity together with whatever extended notions of truth and falsity it was permissible for her to define in terms of them.

Finally, we should note that the radical deflationist's line on foreign utterances can be used to deflect what is apt to seem to be another highly counter-intuitive departure from common sense. Recall that, for me, Radical Deflationism's apparent substitute for 'means,' 'means$_{ss}$,' applies to all and only expressions I understand, and that for example

'Snow is white' means$_{ss}$ that snow is white

is for me analytic, that is, *a priori* and empirically indefeasible. This entails that the displayed meaning$_{ss}$ statement would have been true even if I had used 'Snow is white' the way I actually use 'Coal is black,' and that is apt to seem a capricious departure from our commonsense notion of meaning according to which 'Snow is white' would have meant *that coal is black* if I had used 'Snow is white' the way I actually use 'Coal is black.' While it's true that 'Snow is white' would have meant$_{ss}$ *that snow is white* however I used that sentence, the radical deflationist does have a way of capturing what the ordinary language notion of meaning would say: he can say that in interpreting my utterances in counterfactual worlds where I use 'Snow is white' the way I actually use 'Coal is black,' I wouldn't, and shouldn't, interpret myself

using my egocentric homophonic disquotational notion of meaning$_{ss}$ but would instead interpret myself as though I were speaking a foreign language, in which case I would no doubt use 'Coal is black' to "translate" my counterfactual utterances of 'Snow is white.'

Stage Three: Utterance Understanding

Relative to the pretense that all the sentences of my language are unambiguous and non-indexical, we could view the predicates stipulated to express my egocentric notions of truth and falsity – *viz.* 'true$_{ss}$' and 'false$_{ss}$,' respectively – as applying to *sentences* I understand, and I could represent the radical deflationist as saying that for me to understand a *sentence* is just for it to be, or to be correlated with, a sentence of my mentalese and thereby to have a conceptual role for me. But 'true$_{ss}$' and 'false$_{ss}$' must also have application to *utterances* I understand of sentences I understand, so we need to see what the radical deflationist should say about understanding an *utterance* of a sentence versus understanding the *sentence* uttered. The question presses because it's not sufficient for understanding an utterance of a sentence *S* that one understand *S*. For a start, one can't understand an utterance unless one knows that it occurred. But even if one understands *S* and knows that an utterance of it occurred, that still isn't sufficient for understanding the utterance. For example, the sentence

(1) The force of an allegory's bite can exceed 3,000 pounds

is (we may suppose) an unambiguous eternal sentence I understand (and, of course, know to be false$_{ss}$). Now suppose I hear Fester utter (1). It's very *un*likely that I will interpret his utterance of (1) as a statement that the force of an allegory's bite can exceed 3,000 pounds. If I were wearing my inflationist hat, then how I would interpret Fester's utterance would depend on what I thought his beliefs and intentions were in uttering (1), and no doubt I would conclude that, like Mrs Malaprop, he confuses 'allegory' and 'alligator' and that what he meant in uttering (1) was that the force of an alligator's bite can exceed 3,000 pounds. But whatever interpretation I came up with, I would take that interpretation to be either objectively correct or objectively incorrect,[10] as determined by the intentions and beliefs Fester had in uttering (1). Now that I am trying on the radical deflationist's hat, however, I must recognize that how I interpret Fester's utterance isn't determined by what I think his beliefs and intentions were in uttering (1), for I will recognize that, just as there is no objective fact of the matter as to what his utterance means, so, too, there will be no objective fact of the matter as to what he believes and intends in uttering (1). How I interpret Fester's utterance is simply a matter of which sentence of mine I use to process the utterance, and *that* will be determined by which mapping of his utterance onto a sentence of mine will best satisfy my interests in interpreting him. If, for example, I process the utterance using my sentence

(2) The force of an alligator's bite can exceed 3,000 pounds,

then I am interpreting his utterance as a statement about the force of an alligator's bite. In other words, my saying that Fester's utterance is true$_{ss}$ doesn't mean that it's true on the objectively correct interpretation of it ("the idea of a 'correct understanding' of a sentence or utterance is a semantic notion that has no place when we are discussing purely disquotational truth"; Field, 1986, p. 134); it means that it's *true-as-I-understand-it*, and I'm free to understand it any way I regard as appropriate in the circumstances. Evidently, then, the radical deflationist must say that the sense in which I must understand an utterance – even an utterance of a sentence of my own idiolect – in order for my predicates 'true$_{ss}$' and 'false$_{ss}$' to

have application to it is simply that I process it with a sentence I understand, that is, a sentence that has a conceptual role in my mentalese. Consequently, since I have decided to process Fester's utterance with my sentence (2), then that sentence represents my understanding of the utterance, and when I claim that the utterance is $true_{ss}$ I'm claiming that it's true-as-I-understand-it, which makes that claim cognitively equivalent for me to Fester's utterance of (1), which in turn makes it cognitively equivalent for me to (2). Ways of interpreting utterances even of sentences I understand are in effect ways of translating those utterances into sentences of mine. There is no question of a translation being right or wrong, of being the *correct* or *incorrect* translation of an utterance into my language, only of a translation's being better or worse in a way that is "highly context-dependent (since the purposes for which they are better or worse vary from one context to the next)" (Field, 1994a, p. 128). If I'm unable to interpret an utterance, then "it makes no sense [for me] to inquire whether it is true (except perhaps in counterfactual terms about how I would interpret it under definite conditions)" (Field, 1986, p. 62). The radical deflationist's treatment of utterances of sentences we understand is essentially the same as his treatment of foreign language sentences we don't understand.

Stage Four: Belief

What can Radical Deflationism say about our talk of beliefs being true or false? It would, I believe, be something along the following lines (see Field, 2001b). We start with a stipulative definition:

x believes S* iff *S* is a sentence of *x*'s mentalese that is tokened in *x*'s belief box.[11]

If we were radical inflationists we would then want some refinement of the view that:

- *x* has a belief iff for some proposition *p*, *x* believes *p*
- *x* believes *p* iff, for some *S*, *x* believes* *S* and *S* means *p* in *x*'s mentalese,
  where 'means' expresses a certain use-dependent relation between sentences and propositions, and thus a relation that demands an explication in non-intentional terms.

There are a couple of reasons why the radical deflationist can't say that. One is that it requires a use-dependent notion of meaning and truth-conditions (if *S* means *p* then *S* is true iff *p* is true), what Field calls *correspondence truth-conditions*. The other is that the inflationist account of belief presupposes that there can be a fact of the matter about what someone believes, the content of her belief, even if we are unfamiliar with her and unable to understand the sentences she believes*; but the radical deflationist's unrelativized meaning and truth predicates can be applied only to sentences or utterances one understands.

Now, the only unrelativized radical deflationist truth predicates I have are '$true_{ss}$' and '$false_{ss}$.' What can I say about the truth-conditions of (1)?

(1) Ralph believes that *S*.

Well, if Ralph believes* '*S*' and '*S*' is a sentence I understand, then if I'm going by what '*S*' $means_{ss}$, I can accept (1). But suppose that the sentence Ralph believes* is 'Exercise is enervating' and that he uses that sentence the way I use 'Exercise is energizing.' In that case I should no doubt decide my interests are best served by "translating" 'Exercise is enervating' in Ralph's mentalese as my sentence 'Exercise is energizing.' This is just to repeat what in

effect we already knew – that the radical deflationist should recommend that, if '*S*' is a sentence I understand and I know both that Ralph believes* '*S*' and that Ralph uses '*S*' the same way I do, then I should believe* 'Ralph believes that *S*,' but if I decide '*S*' in Ralph's language is best "translated" as my sentence '*S′*,' then I should believe* 'Ralph believes that *S′*.' In any case, I must recognize that there can be no fact of the matter as to what Ralph believes that would make a belief report about him objectively correct or objectively incorrect, but should recognize that there is nothing to constrain what I take him to believe other than my pragmatic concerns in ascribing beliefs to him. This of course generalizes, *mutatis mutandis*, to every kind of propositional-attitude report.

Again, this is pretty radical stuff, even more radical than the radical deflationist's line on foreign language sentences I don't understand. If I know that Tamiko believes* the Japanese sentence *S* but I don't understand *S*, then, given Radical Deflationism, I must say that there is no fact of the matter as to the truth-conditional content of the belief state Tamiko's believing* *S* realizes. The most I can say is that she has a belief that is true relative to certain ways, and false relative to other ways, of correlating *S* with a sentence of mine. The same is even true of a person who thinks in English when she believes* a sentence and I'm undecided about which sentence of mine I should use to understand that sentence. Since I can be said to understand only an extremely minute fraction of the sentences believed*, desired*, or intended* by all the people there are, were, or will be, we see that Radical Deflationism's line on truth talk about propositional attitudes entails a pretty radical solipsism. For not only must I conclude that there can be no fact of the matter as to whether any native speaker of Japanese ever got married, bought a Honda, or applied to graduate school, I must now conclude that for nearly every person who has ever lived there is no fact of the matter as to whether any of them ever bought anything, accepted a job, or told a lie, or did anything else the doing of which requires acting with particular intentions.

Stage Five: Indexicality and Ambiguity

Up to now I have been pretending that my language is unambiguous and non-indexical. Given that pretense I could say that the egocentric disquotational semantic notions Radical Deflationism licenses for me could be introduced with the stipulations that:

'Means$_{ss}$' applies only to expressions I understand, and every instance of the schema
  '*S*' means$_{ss}$ that *S*
is for me *a priori* and empirically indefeasible.

'True$_{ss}$' and 'false$_{ss}$' apply only to sentences I understand, and every instance of the schemas
  '*S*' is true$_{ss}$ iff *S*
  '*S*' is false$_{ss}$ iff not-*S*
is for me *a priori* and empirically indefeasible, and therefore instances of
  '*S*' is true$_{ss}$
  '*S*' is false$_{ss}$
are cognitively equivalent for me to their corresponding instances of
  S
  Not-*S*.

'*Refers*$_{ss}$' applies only to names I understand, and every instance of the schema
  If *n* exists, then '*n*' refers to *n*

is for me *a priori* and empirically indefeasible.

'True-of$_{ss}$' applies only to predicates I understand, and every instance of the schema
'*F*' is true-of$_{ss}$ *o* iff *o* is *F*
is for me *a priori* and empirically indefeasible.

But what will Radical Deflationism stipulate for my semantic$_{ss}$ expressions when it recognizes that my language is ambiguous and indexical, and therefore, presumably, has to provide completions for forms such as the following?

(1) 'Visiting relatives can be boring' means$_{ss}$ …
(2) 'She is ready' means$_{ss}$ …
(3) 'She' means$_{ss}$ …
(4) A token of 'she' refers$_{ss}$ to *x* iff …
(5) A token of 'She is a novelist' is true$_{ss}$ iff …
(6) 'Tall' means$_{ss}$ …
(7) A token of 'tall' is true-of$_{ss}$ a thing iff …

It's not clear to me that the radical deflationist is able to provide completions of any of these forms that she would find acceptable. Consider (3), (4), and (5). Suppose the pronoun 'she' has only its referential use in my language. Then one sort of inflationist would say (at least to a first approximation):

The meaning of 'she' is that function *f* such that, for any token τ of 'she' and any *x*, $f(\tau) = x$ iff *x* is the female to whom the speaker referred with τ; nothing otherwise.

For any *x* and token τ of 'she,' *x* is the referent of τ iff $f(\tau) = x$; otherwise τ has no referent.

A token τ of 'She is a novelist' is
true iff for some *x*, *x* is the referent of the token of 'she' contained in τ and *x* is a novelist;
false iff for some *x*, *x* is the referent of the token of 'she' contained in τ and *x* is not a novelist;
neither true nor false iff nothing is the referent of the token of 'she' contained in τ.

What might the radical deflationist say?

Field discusses indexicality. He opens that discussion by saying he wants his notion of pure disquotational truth to apply to tokens of indexical sentences but that "one substantial worry about [Radical Deflationism] is whether it can accommodate this."[12] He doesn't say why this is a substantial worry, but he does say why he thinks it shouldn't be any worry at all. Let '$S(i_1, \ldots, i_n)$' be schematic for any sentence of mine with *n* occurrences of indexicals or demonstratives. Field first defines a relativized truth predicate for the sentence-*type* '$S(i_1, \ldots, i_n)$':

'$S(i_1, \ldots, i_n)$' is *true relative to a sequence of objects* $< a_1, \ldots, a_n >$ iff $S(a_1, \ldots, a_n)$.

For example, 'She never loved him' is true relative to < Jill, Jack > iff Jill never loved Jack. That defined notion of truth-relative-to-a-sequence is then in effect used to define the application of 'true$_{ss}$' to tokens of '$S(i_1, \ldots, i_n)$':

(A) A token of '$S(i_1,\ldots, i_n)$' is true$_{ss}$ iff there are objects $a_1, \ldots, a_n$ that I "associate" with the uttered tokens of 'i_1,' …, 'i_n,' respectively, and '$S(i_1,\ldots, i_n)$' is true relative to $< a_1, \ldots, a_n >$.[13]

Field goes on (nearly enough) to explain "association" in a way that implies that I associate objects $a_1, \ldots, a_n$ with the indexicals 'i_1,' …, 'i_n' just in case I processes the utterance of '$S(i_1,\ldots, i_n)$' with a sentence '$S'(\alpha_1, \ldots, \alpha_n)$' of my mentalese such that (i) '$S'(\alpha_1, \ldots, \alpha_n)$' is true$_{ss}$ iff $S'(\alpha_1, \ldots, \alpha_n)$ and (b) $\alpha_1 = a_1, \ldots, \alpha_n = a_n$.[14]

It may be that Field didn't really intend (A) as a proper definition, but merely as an account of how I interpret utterances of indexical sentences when I'm actually party to them, for as a definition, (A) is problematic. Read as a definition, there is prefixed to the definition an implicit necessity operator: 'Necessarily, a token of "$S(i_1,\ldots, i_n)$" is true$_{ss}$ iff ….' But so read, (A) entails that the only tokens of '$S(i_1,\ldots, i_n)$' that can have truth$_{ss}$-values are those that I interpret, and that would mean that if when I'm not around you say 'She was a novelist' intending to communicate that George Eliot was a novelist, then your uttered token of 'she' doesn't refer$_{ss}$ to anyone and your utterance has no truth$_{ss}$-value. But when Field said he wanted his notion of pure disquotational truth – which in my case is 'true$_{ss}$' – to apply to indexical utterances he clearly intended to include utterances in my absence of indexical sentences by those who speak my language. Would Field want to revise (A) to (A′)?

(A′) A token of '$S(i_1,\ldots, i_n)$' is true$_{ss}$ iff there are objects $a_1, \ldots, a_n$ that if I were party to the utterance of that token I would associate with the uttered tokens of 'i_1,' …, 'i_n,' respectively, and '$S(i_1,\ldots, i_n)$' is true relative to $< a_1, \ldots, a_n >$.

I doubt he would want to make that revision. For one thing, counterfactuals like

> If I had been party to α's utterance of 'She's a novelist' I would have associated β with the uttered token of 'she'

rarely have determinate truth-values, and, for another thing, if at the moment of your utterance I were to have a cerebral infarction which causes me to associate your utterance of 'she' with Rosa Luxembourg, would Field really want to say that your utterance is true$_{ss}$ just in case Rosa Luxembourg was a novelist?

I'll skip what the radical deflationist (i.e., Field) has to say about ambiguity, since it runs pretty much along the lines of what he says about indexicality.

Stage Six: Radical Deflationist Truth in Explanation – Where Push Comes to Shove

Radical Deflationism's claim, we already noticed, isn't that the disquotational semantic notions it provides are the notions expressed by ordinary language semantic terms. Its claim is that the semantic notions it provides can do all the explanatory work that semantic notions can legitimately be expected to do. This is where push comes to shove, for the plausibility of Radical Deflationism rides entirely on the plausibility of that claim.

We need to be clear about what the issues really are. No one disputes that semantic notions play an important role in enabling us to acquire knowledge about the world from what people write and say, and in explaining and predicting their behavior. What we want to know is what those roles are and what enables semantic notions to play them. There is also a question about *which* semantic notions we are talking about, and here we find an

assumption that is sometimes made, but is wrong. To see the mistake I have in mind, suppose Jane offers the following two explanations:

(A) I know that Clyde's middle name is 'Ignatz' because he told me it was; he wouldn't have told me that unless he believed he knew it; and his believing that he knows it is extremely good evidence that his belief is *true*.

(B) Of all the many doctors Frank consulted, Dr Jones was the only one who succeeded in relieving his symptoms, and that was because she was the only doctor whose belief about the cause of those symptoms was *true*.

Now suppose a radical deflationist were to argue that (i) egocentric disquotational truth can play the role truth plays in (A) and in (B), for (ii) the explanation (A) offers remains the same if we replace the sentence

his believing that he knows it is extremely good evidence that his belief is true

with the sentence

his believing that he knows that his middle name is 'Ignatz' is extremely good evidence that his middle name is 'Ignatz';

and the explanation (B) offers remains the same if, using substitutional quantification, we replace the phrase

whose belief about the cause of those symptoms was true

with the phrase

whose belief about the cause of those symptoms was such that ΠS(('S' purports to identify the cause of Frank's symptoms & Dr Jones believes that S) → S).

The problem with this argument is that (ii) doesn't entail (i). The reason it doesn't is that *it doesn't address the truth-conditions entailed by the that-clauses in the explanations' belief ascriptions*. The heavy lifting in propositional-attitude explanations is done by the *contents* ascribed to the beliefs (and other propositional attitudes) cited in those explanations, and, as we have already observed, those contents are truth-condition entailing: the claim that Clyde believes *that his middle name is 'Ignatz'* entails that Clyde has a belief that is *true if and only if his middle name is 'Ignatz'*. The difficult task for the radical deflationist isn't to explain overt uses of 'true' like those in (A) and (B); it's to explain the use of *propositional content* in those and other propositional-attitude explanations.

The notions expressed by our ordinary language propositional-attitude expressions – 'believes that such and such,' 'desires such and such,' 'intends to do such and such' – do a lot of extremely important explanatory work for us. The plausibility of Radical Deflationism turns on its ability to do the same work. Before we address that question directly we should first say something about some of the key features of our propositional-attitude expressions as we actually use them. Here is a brief annotated inventory of what some of those features appear to be.

a) Propositional-attitude predicates such as 'believes that Brutus killed Caesar' and 'intends to buy a villa in Cap Ferrat' apply interpersonally and univocally, even among monolingual speakers of different languages. We depend on this as when we say things like 'The vast majority of citizens of other countries believe that the United States had no legitimate reason for invading Iraq.' The same is true of predicates that entail propositional-attitude properties, such as 'bought a villa in Cap Ferrat,' which apply to a person only if she acted with certain intentions.

b) Do propositional-attitude properties play a causal-explanatory role in propositional-attitude explanations, as, for example, when one explains that Ava passed the salt to you because she believed that you asked her to? Define 'causal-explanatory.' If Henry took pain medication because he was having a migraine, is the property of having a migraine playing a "causal-explanatory role" in that explanation? If you say yes, I have no trouble with that: I assume you say it because Henry's migraine was a cause of his taking the pain medication and that the cause's having the property of being a migraine is a causally relevant property of it, in that, all else being equal, the cause wouldn't have been a cause had it not had that property. But if you say that being a migraine plays a "causal-explanatory" role in the explanation of Henry's taking pain medication, then parity of reasoning should require you to say that the property of being a belief *that you asked her to pass the salt* is playing a causal-explanatory role in the explanation of Ava's passing you the salt. It's less clear what you would mean if you said that the property of having a migraine isn't playing a causal-explanatory role in the Henry explanation. I might guess that you accepted something like Jaegwon Kim's "principle of explanatory exclusion," that two events can't have two separate and complete explanations (see, e.g., Kim, 1988) and that you thought the migraine explanation couldn't be identified with a more basic non-psychological explanation. In that case I would want to have a conversation with you about explanatory exclusion. The important question, of course, is what radical deflationists mean by 'causal-explanatory role.' I'm not aware of any place where Field tries to say what it is. Stephen Leeds implies that a notion plays a causal-explanatory role if it's needed in a causal law (Leeds, 1978, p. 116), but I should think that if there are laws of our commonsense propositional-attitude psychology, then they are content-involving laws such as 'For any person x and state of affairs s, if x desires s to obtain and, for some act *A*, believes *that s will obtain only if x does A*, and if such-and-such other conditions are satisfied, then, ceteris paribus, x does *A*.'[15] The notion of a causal-explanatory role (whatever exactly that notion is) is important in the present discussion because the radical deflationist claims both that we need inflationist semantic notions only if we need semantic notions that are explicable in physical terms, and that they need to be so explicated if, but only if, they are needed to play a causal-explanatory role.

c) It's pretty clear that the properties expressed by the propositional-attitude predicates we actually use are use-dependent contingent properties. But are they physicalistically explicable? Define 'physicalistically explicable.' Field, who makes heavy use of the expression, couldn't define it when he wrote his Tarski paper. There he is deliberately vague as to what it would be for a notion to be explicable in physical terms. He says it's "very hard to take seriously" the popular idea that explicability in terms of physical facts requires that "for every acceptable predicate '$P(x)$' there is a formula '$B(x)$' containing only terminology from physics, such that '$\forall x(P(x) \Leftrightarrow B(x))$' is true" (Field, 1972, p. 11, fn. 9), but he admits not to having a precise characterization of his own and pretty much concludes that the notion is whatever scientific practice needs it to be. But in a paper published 20 years after his Tarski paper, he suggests that "if we are to accept a special-science explanation of something, we are

committed to the possibility in principle of finding a physical explanation of that thing *in which the structure of the special-science explanation of it is preserved*" (Field, 1992, p. 278; emphasis is Field's). This strikes me as implausible: 'Jack and Jill adopted a child because they wanted very much to have a child but learned that it was impossible for them to conceive one' might sketch a correct propositional-attitude explanation of the fact that Jack and Jill adopted a child, but what on earth would a physical explanation of that fact look like? I do think that if a propositional-attitude explanation E is correct it must in principle be possible to explain why E is correct in terms of underlying physical facts and their relation to the propositional-attitude facts E entails, where such an explanation would consist primarily in explaining why the propositional-attitude causes postulated by E are causes, and why the propositional-attitude properties of those causes are essential to their being causes. Such an explanation would no doubt necessitate myriad physical explanations of myriad and sundry physical facts, but I very much doubt that any one of those explanations would preserve E's structure. I don't take the fact that correct propositional-attitude explanations aren't shadowed by structurally identical physical explanations to show that propositional-attitude properties aren't reducible to physical properties in the sense that, for every propositional-attitude property Φ there is a logically complex property Π built up from physical properties such that having Π is metaphysically necessary and sufficient for having (some suitable precisification of) Φ. Field has said that one reason he came to doubt the correspondence theory of his Tarski paper was that it "proved extraordinarily difficult to develop the details of an adequate correspondence theory" (Field, 1986, p. 67). This is misleading. What has proved difficult is getting a correspondence theory using only the tools available to a philosopher without having to leave his armchair. Why should we expect properties that are metaphysically equivalent to our propositional-attitude properties to be discoverable by the slow-moving and storage-limited information-processing of our brains?

So much for ground clearing. It is important to keep in mind that the question

> Are semantic notions needed to play a causal-explanatory role in psychological or linguistic explanations?

isn't the same as the question

> Can the semantic notions provided by Radical Disquotationalism do the explanatory work that is done by the semantic notions we actually use?

For all we yet know, the answer to both questions is no; for whether or not the propositional-attitude and other semantic properties expressed by our propositional-attitude and semantic expressions *as we actually use them* are "causal-explanatory" or "physicalistically explicable" properties, they are inflationist properties in that they have application to all people, regardless of the languages they speak, and are use-dependent contingent properties. It's therefore crucial for the radical deflationist to show that in some relevant sense the explanatory benefits we get from the propositional-attitude notions we actually use can be had by using the notions he would substitute for them. Here is one reason that may not be such an easy thing to show. Common sense supposes that there are billions of intentional actions performed every day, nearly all of them by people with whom I'm unacquainted, speaking languages I don't understand, and that nearly all of these actions enjoy correct propositional-attitude explanations. For example, common sense has no trouble allowing

that the following explanation might be true, and even that I might know it, notwithstanding that Olga is a monolingual speaker of Russian and I don't understand a word of that language other than 'da' and 'nyet':

(E) Olga has petitioned to have her marriage annulled because she learned that her husband was married to someone else when he married Olga.

Explanations like (E) are extremely useful to us in myriad ways, but according to Radical Deflationism, I can't allow that (E) is correct. In fact, if I accept Radical Deflationism, I must conclude that the vast majority of propositional-attitude explanations thought to be correct are attempts to explain facts that don't exist in terms of other facts that don't exist. For I must suppose that the only people for whom there may be a fact of the matter as to the truth-conditional contents of their propositional attitudes are myself and the small number of people whose utterances I can understand – and then 'truth' in 'truth-conditional content' must be 'truth$_{ss}$' and 'fact' in 'fact of the matter' must be 'fact$_{ss}$,' the very thin notion of fact that goes with truth$_{ss}$. As for the billions of other propositional attitudes, I must recognize that they can have truth-conditional content only relative to their being correlated in this, that, or the other one of infinitely many ways with my sentences, where no one of those ways is the correct way of correlating them with my sentences. That is why I would have to conclude that the vast majority of propositional-attitude explanations thought to be correct are attempting to explain what they are attempting to explain by citing non-existent facts; at the same time, in most cases the facts they are attempting to explain would also be non-existent, because they would be facts, such as the fact that Olga petitioned to have her marriage annulled, about intentional actions that can be performed only by someone with beliefs and intentions with certain contents.

It might be thought that, while Radical Deflationism precludes me from accepting (E), there is something pretty close to (E) that I can accept, and which will serve me pretty much as well as (E) does. Wearing my commonsense hat, I have no trouble believing that Olga came to believe that her husband was already married to someone else when he married Olga, that her coming to believe that led to her intending to petition to have her marriage annulled, and that that led to her so petitioning. Wearing my radical deflationist hat, I can't, for example, say that Olga believes that her husband was already married to someone else when he married Olga, because according to Radical Deflationism my sentence

Olga believes that her husband was married to someone else when he married Olga

is equivalent to

Olga believes* a sentence of her Russian idiolect that is true$_{ss}$ – that is, true *as I understand it* – iff Olga's husband was married to someone other than Olga when he married Olga.

But for a sentence of Olga's Russian idiolect to be true as I understand it I must understand it, and in this case that would require me to have decided to interpret her sentence as equivalent to a certain sentence of mine. Of course, I do know that there is a standard way of translating Russian into English – no doubt a way of translating constructed by highly educated and linguistically sophisticated people who are completely bilingual in English and Russian – and I might have it on good authority that that method of translation would

correlate the sentence Olga believes* with my sentence 'Olga's husband was married to someone other than Olga when he married Olga,' and that, consequently, while I can't accept (E), I can accept something along the lines of

(E*) (E) is true relative to the standard way of translating Russian into English,

and if I'm determined to interpret Olga's sentences as equivalent to the English sentences with which they are correlated by the standard way of translating Russian into English, then I shouldn't be worse off than if I could accept (E) directly. There are, however, two problems with this response. The first is that it's question-begging for the radical deflationist to assume he can rely on the standard ways of translating foreign languages into English, because if those translation schemes were constructed on the basis of an inflationist notion of sameness of meaning, then the radical deflationist would find that the benefits of using his theory-approved notions were due to their unwitting reliance on an inflationist conception of semantic notions. For the radical deflationist, a translation scheme can be selected for interpreting foreign sentences only if it's selected on the basis of radical-deflationist-approved properties of sentences, such as conceptual-role and indication-relation properties, which are as available to the radical deflationist as they are to the inflationist. An inflationist who accepts the hypothesis that we think in a language-like system of mental representation will say that

x believes *p* iff, for some sentence *S* of *x*'s mentalese, *x* believes* *S* and *S* means *p* in *x*'s mentalese,

and she will assume that which proposition a mentalese sentence means is determined by some yet unknown package of properties of *S* that will include conceptual-role and causal-relation properties that may explain, *inter alia*, why *x*'s believing *p* is to some degree evidence that *p* is true. For the inflationist, the explanatory role that *x*'s believing *p* is able to play rides piggy-back on the explanatory role played at another level of explanation by the meaning-determining package of properties of the sentence *S* such that *x*'s believing* *S* realizes *x*'s believing *p*. The radical deflationist can agree with the inflationist that it's a conceptual-role/causal-relation package of the sentence *x* believes* that is really carrying the explanatory load without agreeing with her that that package is able to define that relation which a sentence in a person's mentalese must bear to a proposition in order for it to mean that proposition in that person's mentalese. Consequently, when he needs to interpret the sentences Olga believes*, desires*, or intends* he will look for a correlation scheme that enables him to exploit the explanatorily relevant properties of those sentences.

That correlation scheme may not be so easy to find. For one thing, how will the radical deflationist know *which* conceptual-role and causal-relation properties are the ones doing the underlying explanatory work? In fact, he can't know. In order merely to know what the explanatorily relevant conceptual-role properties are would require a complete computational psychology, and we are years from having that. As for picking out the causal properties one would need to explain why certain of a person's beliefs reliably indicate their truth, just consider how much neurophysiology one would have to master even to begin making a dent in that task, not to mention that much of what one would need to know about the neurophysiological workings of the brain and central nervous system is nowhere near being known. An enormous advantage of our ability to give commonsense propositional-attitude explanations and predictions is that we can give them without knowing anything about the

underlying properties that make those explanations and predictions possible, just as a person who knows nothing about how his computer works, or even the programs it's running, can explain and predict its behavior. Nor can the radical deflationist say with any degree of aplomb that he will select a system of correlation that correlates a sentence *S* of Olga's with a sentence *S′* of his only if *S* and *S′* share more or less the same relevant conceptual-role and causal-relation properties, for how will he know which properties are relevant and how will he know that the conceptual-role and causal-relation properties that play a certain role in the explanation and prediction of Olga's behavior play the same sort of role in the explanation and prediction of his relevantly similar behavior?

So far I've been wondering what sort of an explanation, if any, of Olga's petitioning for an annulment Radical Deflationism makes available to me if I don't understand the sentences Olga believes*, desires*, or intends*. But the most serious problem for the theory concerns how I might explain Olga's behavior when I *do* understand those sentences. To see what I mean, let's return to

(E) Olga has petitioned to have her marriage annulled because she learned that her husband was married to someone else when he married Olga

and let's suppose that I can understand Olga's sentences so that I have no trouble accepting as true_{ss} such sentences as 'Olga believed that her husband was married to someone else when he married Olga,' 'Olga intended to petition for an annulment,' and 'Olga has petitioned for an annulment.' Then it will be the radical deflationist's claim that he can accept (E) *when it's read in the way his theory requires him to read it*, thereby demonstrating that the only semantic notions we need in order to give such explanations as (E) are the ones his theory makes available. That claim, however, is false: the radical deflationist cannot accept (E) when the content-involving notions in it are taken to be his theory-approved notions. For (E) implies (a) that Olga's coming to believe that her husband was married to someone else when he married Olga was a cause of Olga's petitioning for an annulment and (b) that Olga's coming to have that belief was a cause of her petitioning for an annulment *because* it was a belief *that her husband was married to someone else when he married Olga*, where this implies that, all else being equal, she would not have petitioned for an annulment if her belief hadn't had that content. It ought to be clear that (b) is every bit as important as (a): without (b) we should have no explanation of why Olga's coming to have a certain belief was a cause of her petitioning for an annulment, and (b) helps to explain how someone who knew *that Olga learned that her husband was married to someone else when he married Olga* might have predicted that she was likely to seek an annulment. When, however, we read (E) in the way the radical deflationist requires it to be read, we see that, while he can accept (a) he can't accept (b). For suppose the radical deflationist in question is myself. Then I can't accept (b) because I must say that

Olga believes that her husband was married to someone else when he married Olga

is equivalent to

Olga believes* a sentence that is true_{ss} – that is, true as *I* understand it – iff 'Olga's husband was married to someone else when he married Olga' is true_{ss},

but it's plainly false that she wouldn't have petitioned for an annulment if the explainer hadn't interpreted a sentence of Olga's in a certain way.

Field is sensitive to this problem, but nowhere does he explicitly offer his official response to it. He acknowledges that "the most serious worry about [Radical Deflationism] is that it can't make sense of the explanatory role of truth conditions … in explaining behavior," but goes on to say that "unfortunately it is a big job even to state the worry clearly, and a bigger job to answer it; I must save this for another occasion" (Field, 1994a, p. 127). Apart from some remarks in the postscripts to "Mental representation" and "Deflationist views of meaning and content" in *Truth and the Absence of Fact* (Field, 2001a), the job is still being saved. But in the postscript to "Mental representation" he does suggest a response he may wish to develop. There he suggests that for the radical deflationist propositional-attitude explanation might be what he calls "projective" explanation. This alludes to the simulation theory of psychological explanation first proposed by Robert Gordon and further developed by Gordon, Alvin Goldman, and others (see, e.g., Gordon, 1986, and Goldman, 1989). In a projective explanation I explain another's behavior not by subsuming it under the generalizations of a folk theory that applies equally to all folk, but by imaginatively putting myself in the other's shoes and asking what I would do if I were in them. Field's gloss on this is that:

> [A "projective" explanation of someone's behavior involves] reference to the explainer's language even though that is not causally relevant to the behavior. When explaining a person's behavior (say the raising of his gun) in terms of his belief that there is a rabbit nearby, what I am in effect doing is explaining the behavior in terms of his believing* a representation that plays a role in his psychology rather similar to the role that 'There are rabbits nearby' plays in mine.… Such an explanation is still basically non-intentional: truth conditions play no real explanatory role. Of course, there is a sense in which my *sentence* 'There are rabbits' plays an explanatory role here: obviously not as a causal factor in the explanation, but as a device we use in picking out the agent's internal representation (which is a causal factor). (Field, 1978, p. 78)

As regards (E), what this projectivist line suggests is that the radical deflationist should reject (E), which does give an explanatory role to truth-conditions (whether or not that role is a causal-explanatory role), and, imagining Field to be the explainer, replace it with:

(E′) Olga has petitioned to have her marriage annulled and a cause of that petitioning is that Olga believed* a sentence that plays a role in her psychology that is similar to the role 'Olga's husband was married to someone else when he married Olga' plays in Hartry Field's psychology.

I know Hartry Field pretty well, but I'm very reluctant to say I know the role that the sentence about Olga's husband plays in his psychology. But there is a bigger problem. A monolingual speaker of Hindi could accept the explanation expressed by (E′) without having any idea of how to translate Field's sentence about Olga's husband into *her* language. As Field happily acknowledges, the meaning or truth-conditions of 'Olga's husband was married to someone else when he married Olga' plays no role in the explanation (E′) communicates. (E′) gives no more of an explanation of Olga's petitioning for an annulment than does:

> Olga has petitioned to have her marriage annulled and a cause of that petitioning is that Olga believed* a sentence that plays a role in her psychology that is similar to the role a certain sentence plays in Hartry Field's psychology.

The use of 'true' as a device for expressing certain infinite conjunctions and disjunctions is a red herring. The most important case for inflationist semantic notions is that such notions go hand in hand with inflationist propositional-attitude notions, and the most important case for them is that they seem to be needed to account for the role those notions play in commonsense propositional-attitude explanations of behavior, such as the fact that Olga petitioned for an annulment *because she believed that her husband was married to someone else when he married Olga*. It's this role that Radical Deflationism seems unable to accommodate.

II

The classical notion of a *proposition* in analytical philosophy is the one first clearly articulated by Frege – namely, that a proposition is an abstract, mind- and language-independent entity that has truth-conditions, and has its truth-conditions both essentially and absolutely (i.e., without relativization to anything).[16] To say that every proposition has truth-conditions is not to say that every instance of the schema

The proposition that *S* is true iff *S*

is true. Frege himself would say that the proposition *that the present King of France is bald*, as well as the sentence 'The present King of France is bald' which expresses it, is neither true nor false owing to the fact that there is no such person as the present King of France. Consequently, he would say that

The proposition *that the present King of France is bald* is true iff the present King of France is bald

is not true, because its left-hand side is false but its right-hand side is neither true nor false. The classical conception of a proposition does however entail that, whatever form a proposition's truth-conditions take, it will be a necessary, and therefore use-independent, fact that it has those truth-conditions. If, *pace* Frege, every instance of the schema

The proposition that *S* is true iff *S*

which expresses a proposition is true, then those instances are necessary, and therefore use-independent, truths. It's also arguable – and has been argued at length by Paul Horwich (1998) – that anyone who possesses the concepts required to understand a true instance of the schema would *ipso facto* implicitly know that it's true.

The preceding paragraph does *not* express a *deflationist* conception of propositional truth; it is what inflationists who use the classical notion of a proposition must say, and there is no "inflationist correlate" that the view seeks to deflate. Hartry Field formulates his deflationist theory without reference to propositions, but he acknowledges that there is a version of his theory that does refer to them (Field, 2015). That version would say that each of

(1) (Π*S*)('*S*' means the proposition that *S*)
(2) (Π*S*)(the proposition that *S* is true iff *S*)
(3) A sentence is true iff the proposition it means is true

is a use-independent conceptual truth, and from that it would follow that

(4) (Π*S*)('*S*' is true iff *S*)

is also a use-independent conceptual truth. Had Field stated his theory this way, his theory would be deflationist by virtue of holding that (1) was a use-independent conceptual truth. No theory is deflationist just by virtue of holding that (2) and (3) are use-independent conceptual truths.[17] The theories of propositional truth of Frege, Frank Ramsey, and Paul Horwich (Frege, 1918; Ramsey, 1927; Horwich, 1998) – as well as the prosentential theory of Dorothy Grover, Joseph Camp, and Nuel Belnap (Grover, Camp, and Belnap, 1975), which is a theory of 'true' as applied both to propositions and to sentences – all hold that (2) and (3) are use-independent conceptual truths, but they also hold that, since use determines meaning, (1), and with it (4), is a use-dependent contingent truth. Yet the literature labels these theories 'deflationist.' I have explained why no theory should be called deflationist *just* on the basis of holding that (2) and (3) are use-independent conceptual truths, but it may be that those theories say other things that warrant thinking of them as having inflationist correlates which they are out to deflate. Before saying another word, however, I must emphasize that I don't see that any issue worth arguing about turns on how anyone uses the word 'deflationist' (or 'minimalist'), and I do not intend in any way to disparage a theory by questioning whether it should be called deflationist. I do, however, want my use of the label to be clear.

Theories of propositional truth which are referred to as deflationist despite their holding that use determines meaning make claims in addition to those which don't on their own warrant pinning the label 'deflationist' on them, and some of these claims arguably do raise important issues possible responses to which invite an inflationist/deflationist division. I will close this chapter by briefly commenting on what I take to be the most important such issue: the issue of *truth-aptness*.

A sentence is *truth-apt* just in case it has truth-conditions (and is therefore apt for being true).[18] Truth-aptness raises the question: What must be the case in order for a sentence to be truth-apt? The question is important because certain prominent answers to it are in conflict with certain theories of the semantics of certain kinds of sentences, such as ethical sentences and indicative conditionals. The issue is of considerable concern to Field. On the one hand, his disquotationalism entails that a sentence *S* is truth-apt just in case

'*S*' is true iff *S*

is syntactically well formed, and, since

- 'Torture is wrong' is true iff torture is wrong
- 'The execution will be held indoors if it rains' is true iff the execution will be held indoors if it rains

are syntactically well formed, Field is committed to saying that ethical sentences and indicative-conditional sentences are truth-apt. On the other hand, he is sensitive to the appeal of theories which deny that such sentences are truth-apt. Field's article "Disquotational truth and factually defective discourse" (Field, 1994b/2001a) is an attempt to show that he can capture what he likes about the theories to which he is attracted in ways that are

compatible with his Radical Deflationism.[19] Frank Jackson, Graham Oppy, and Michael Smith are considerably less sanguine.

The radical deflationist's criterion for truth-aptness is an example of what Jackson, Oppy, and Smith call *syntacticism*. Syntacticism holds that:

> Truth-aptness is purely and simply a matter of syntax.… Provided the sentence can be significantly embedded in suitable constructions – for example, negation, conditional, propositional attitude and truth ascriptions – then, according to the syntacticist, that sentence is truth-apt. (Jackson, Oppy, and Smith, 1994, p. 291)

They are summarily dismissive of Field's view that truth is a property that a sentence has or fails to have independently of the way that the sentence is used by speakers, and their objection to syntacticism is that it's incompatible with the platitude that use determines meaning. More specifically, they argue that (i) a sentence *S* is truth-apt only if there is some set of truth-conditions *C* such that *S* has *C*, but (ii) there is no set *C* of truth-conditions such that a sentence has *C* just by virtue of having certain syntactical properties; truth-conditions can be determined only by "a rich enough pattern of usage" (Jackson, Oppy, and Smith, 1994, p. 293). For theorists who hold that a sentence's truth-conditions are just those of the proposition it expresses, the question

> For any sentence *S*, what must be the case in order for *S* to be truth-apt?

reduces to the question

> For any sentence *S* and proposition *p*, what must be the case in order for *S* to mean *p*?

The theories of propositional truth that are both labeled 'deflationist' and hold that use determines meaning may nevertheless hold accounts of truth-aptness that Jackson *et al.* would say are minimalist (= deflationist) accounts of truth-aptness by virtue of how little those accounts require in order for there to be a proposition that a sentence means. The only theory of this kind they discuss is one they say is held by Paul Boghossian and Crispin Wright (Boghossian, 1990; Wright, 1992) and which they call *disciplined syntacticalism*. For a sentence to be truth-apt according to disciplined syntacticalism it must have the syntactical properties required by syntacticalism, but in addition to that, they quote Boghossian as saying, the sentence's use "must be appropriately disciplined by norms of correct utterance" (Boghossian, 1990, p. 163). Jackson *et al.* call this a "minimalist" account of truth-aptness because, they claim, ethical sentences and indicative conditionals will pass its test for being truth-apt.[20] An objection one might raise to disciplined syntacticalism is that it doesn't say enough to enable us to see how standards of correct utterance determine *which* proposition a sentence means. The objection Jackson *et al.* raise to disciplined syntacticalism

> is not to what it says, but to what it fails to say. It does not say enough about a central platitude governing truth-aptness and its connection with belief.… Our contention is that there is an analytical tie between *truth-aptness* and belief, specifically, the belief of someone who asserts the truth-apt sentence. Part of the story about rich patterns of usage required to confer truth conditions must be a story about using the sentence to express belief. (Jackson, Oppy, and Smith, 1994, p. 294)

This is an objection to disciplined syntacticism because, they claim, it's possible for a sentence to pass the disciplined syntacticist's test for truth-aptness but fail to meet the expression-of-belief criterion. For in order for a state to be a belief – in the sense of 'belief' they intend (they don't deny that there is a sense in which one can say 'I believe that torture is wrong') – it must have a certain functional role, and it's not clear that the state expressed by an utterance of, say, 'Torture is wrong' has that functional role. One reason that isn't clear (they mention others) is that two views, each of which enjoys a positive degree of plausibility, are incompatible with one another. One view is the "Humean doctrine … that no belief can have a conceptual connection with motivation"; the other is that the state we would express in uttering an ethical sentence does have a conceptual connection with motivation: it's arguable, for example, that the state expressed by an utterance of 'Torture is wrong' has "a conceptual connection with being motived against torture" (Jackson, Oppy, and Smith, 1994, p. 298; see also Smith, 2016).

The disciplined syntacticist may not be without a response. Jackson *et al.* claim that it's possible for a sentence to satisfy the criterion the disciplined syntacticist says is sufficient for a sentence's being truth-apt even though uttering the sentence can't express a state with the belief-making functional role. The best strategy for the disciplined syntacticist is not to deny the functional-role criterion for being truth-apt but to argue, first, that satisfying her truth-aptness criterion entails satisfying the functional-role criterion, and, second, that a sentence's satisfying the functional-role criterion is compatible with the sentence's having the features that motivate denying that the sentence is truth-apt. For example, as regards the second part of the best strategy in the case of ethical sentences, the disciplined syntacticist might point out that Hume's doctrine isn't inviolate: nothing precludes a state from having a functional role that makes the state both a belief and a state that one can't be in unless one is also in a certain conative state; nothing, so to say, precludes 'Torture is wrong' from having a conceptual role that precludes the sentence from being tokened in one's belief box unless the sentence 'No one tortures' is tokened in one's desire box. After all, can one believe that one will now leave one's office to teach a class without intending to leave one's office to teach a class? Can one believe that one has an itch without having any inclination to scratch?[21] Now, if it can be shown that, for some proposition *p*, sentence *S* means *p*, then it should be easy to show that one who has command of *S* can be in a state that has whatever functional role is required for its being a belief that *p*. The difficult task for the disciplined syntacticist is giving *the right sort of* account of the nature of propositions generally and of the relation that must obtain between a sentence and a proposition *p* in order for the sentence, or an utterance of it, to express *p*, where the right sort of account is one that explains how little is required by way of the use of a sentence *S* in order to validate the "something-from-nothing" inference from '*S*' to 'The proposition that *S* is true,' and, correlatively, how little is required by way of the use of a predicate *F* in order to validate the "something-from-nothing" inference from '*n* is *F*' to '*n* has the property of being *F*.'[22] It remains to be seen whether the disciplined syntacticist can meet this challenge, but if she can meet it, then showing that it's met might well reveal that her position isn't very far removed from the spirit, if not the letter, of Hartry Field's deflationism.[23]

Notes

1 I use 'deflationist' where others would use 'deflationary' or 'minimalist.'

2 Field (1986; 1994a, reprinted with a new postscript in 2001a; 1994b; 2001b; 2015). Proponents of Radical Inflationism include David Armstrong, Michael Devitt, Paul Grice, Jerry Fodor, David

Lewis, Stephen Neale, Nathan Salmon, Scott Soames, Robert Stalnaker, and the time-slice of Hartry Field who wrote "Tarski's Theory of Truth."

3 'Eternal sentence' is Quine's expression for sentences whose truth-values must remain the same no matter when, where, or by whom they are uttered – for example, 'Bernard J. Ortcutt owes W. V. Quine ten dollars on July 15, 1968,' where 'owes' is stipulated to be tenseless. See Quine (1970, p. 13).

4 The radical deflationist may prefer explicit definitions of 'true_{ss}' and 'false_{ss}' that entail that they apply only to sentences I understand and that make every instance of '*S* is true_{ss}' cognitively equivalent for me to the corresponding instance of '*S*' and make every instance of '*S* is false_{ss}' cognitively equivalent for me to the corresponding instance of 'Not-*S*.' Field in effect suggests that, if we have "a theory of substitutional quantification that avoids the semantic paradoxes," then such definitions can be had by letting 'S' range over unambiguous eternal sentences I understand and then stipulating:

- *S* is true_{ss} iff $\Pi S'((S = \text{`}S'\text{'}) \rightarrow S')$
- *S* is false_{ss} iff $\Pi S'((S = \text{`}S'\text{'}) \rightarrow \neg S')$

(Field, 1994a, p. 120, fn. 17). Definitions in the same vein would also be available for the other semantic_{ss} predicates.

5 Consequently, Field mischaracterizes his deflationism when he says that "a deflationist thinks that a homophony condition guarantees that we are speaking English rather than English*" (Field, 1994a, p. 126).

6 Leeds (1978). Field emphasizes this in all the works cited in n. 2. He qualifies his claim that we need a notion of truth as a device for expressing infinite conjunctions and disjunctions by noting that substitutional quantification affords another way of expressing them (Field, 1994a, p. 120, fn. 17).

7 I am only considering how the radical deflationist might legitimate talk of the truth or falsity of foreign sentences one doesn't understand. Field claims that for foreign sentences one does understand one may simply apply one's own egocentric predicate, as it applies to every sentence one does understand. So, for example, if 'Der Schnee ist weiss' is an unambiguous eternal sentence I understand, I can say:

'Der Schnee ist weiss' is true_{ss} iff der Schnee ist weiss.

8 An indication relation is a relation between a sentence and external state-of-affairs which makes believing or asserting the sentence a reliable indication that the state-of-affairs obtains.

9 Field (1986, p. 62). Because I accept the deliverances of Google Translate, I interpret the Portuguese sentence 'Al e Betty dançou o macarena' as being equivalent to my sentence 'Al and Betty danced the macarena' – that is how I choose to interpret the Portuguese sentence. There is some unclarity in Field as to whether I can say that 'Al e Betty dançou o macarena' is true_{ss} iff Al and Betty danced the macarena or whether the most I can say is that the Portuguese sentence is true relative to Google Translate's correlation scheme iff Al and Betty danced the macarena. I have decided to use the first option throughout: otherwise (a) Field's gloss of 'pure disquotational truth' as 'true as I understand it' becomes confusing (I *understand* 'Al e Betty dançou o macarena' as equivalent to 'Al and Betty danced the macarena'); (b) it's impossible to make sense of Field's allowing that 'Der Schnee ist weiss' is for him true in the pure disquotational sense iff der Schnee ist weiss; (c) he certainly doesn't want to say that utterances of ambiguous or indexical sentences can't be true in the pure disquotational sense, but they, too, require deciding on a way of mapping such utterances onto sentences of one's mentalese; and (d) allowing that foreign sentences may be true_{ss} in this way lessens a little some of the counter-intuitive consequences of Field's theory. At the same time, I don't see that anything important turns on which of the two options one uses, since each is easily translated into the other.

10 Or I would take it to be indeterminate whether it was objectively correct or objectively incorrect, or indeterminate whether it's indeterminate whether …– but let's not get into that!

11 For a sentence to be tokened in a person's "belief box" is for it to be tokened in a state that has the functional property responsible for a state's being a belief. See Schiffer (1981).

12 Field (1994a, p. 134). Field uses 'utterance' where I use 'token'; I prefer 'token' because accounts of reference in terms of tokens lend themselves to slightly simpler formulations.

13 Field (1994a, p. 136). Field says that typically the objects he associates with the indexicals will be ones he takes "the producer of the utterance to have intended" (p. 136, fn. 31). This is somewhat misleading, however, since for Field, as we have seen, there is no objective fact as to what objects the producer intended: from Field's perspective, what intentions a person has is determined by how he, Field, decides to interpret the sentences the producer intends*.

14 Field (1994a, p. 136). Field's idea here is not that he will have a name or definite description to replace every indexical or demonstrative, but that uttered indexicals or demonstratives may be thought of as represented by subscripted mentalese indexicals or demonstratives that arise on the spot and function in processing like names.

15 My own view (1991) is that there aren't folk psychological, or even special-science, laws because if there were such laws they would have to be "*ceteris paribus*" laws, and I doubt that generalizations needing *ceteris paribus* clauses are capable of stating laws. For an opposing view, see Fodor (1991).

16 Frege (1892) and (1918); Frege called propositions 'thoughts.'

17 This puts me in agreement with Field's remark at the beginning of (1986) about where "the real philosophical problem lies."

18 Strictly speaking, we should say that a sentence is *truth-apt* just in case it's possible for tokens of it to have truth-conditions, for few of the sentence-types we utter have truth-conditions. This is obvious as regards indexical and ambiguous sentences, but it also applies to vague sentences, and virtually every sentence is vague to at least some extent. Vague sentence-types don't have truth-conditions because of a phenomenon we may call *penumbral shift*. A vague term's *penumbra* is that area of logical space wherein the term's application is anything other than unqualifiedly determinately correct or unqualifiedly determinately incorrect. The penumbra of a vague term may shift somewhat from one context of utterance to the next, and these shifts entail shifts in the term's application conditions, so that, for example, 'bald' is true of Harry in one context of utterance, false of him in another, and neither determinately true nor determinately false of him in still another context of utterance. In this way, tokens of the sentence 'Harry is bald' produced in those three contexts will have somewhat different truth-conditions from the other two tokens. Having said that, however, I'm going to ignore it and join Jackson *et al*. in simply speaking of a sentence's truth-conditions.

19 The article also discusses vague discourse and discourse involving referential indeterminacy as other kinds of "factually defective" discourses that might appear to be in conflict with Radical Disquotationalism.

20 It's actually not clear that ethical sentences will be truth-apt on such expressivist theories as emotivism or R. M. Hare's prescriptivism (see, e.g., Hare, 1952). For on these theories, there are no norms of correct utterance to legislate between competing ethical claims, nothing one can say to the Nazi who says 'Jews ought to be exterminated' to prove to him that he is wrong. Jackson *et al*. do say that "some might quarrel with the claim that the [norms of correct utterance] are firm enough in the ethical case," but the "some who might quarrel" can't be emotivists or prescriptivists, for their position is quite clear that on their views no ethical judgment can be objectively correct or incorrect (when one protests to the Nazi that his claim isn't correct, all one is doing is countering one emotive or prescriptive utterance with another).

21 See Schiffer (2003, ch. 6), Michael Smith's (2016), and Schiffer (2016), which is a response to Smith.

22 In (Schiffer, 2003, p. 61) I explained that we have a something-from-nothing inference "when from a statement involving no reference to an *F* we can deduce a statement that does refer to an

F …. [For example,] from the statement 'Lassie is a dog', whose only singular term is 'Lassie', we can validly infer the pleonastic equivalent 'Lassie has the property of being a dog', which contains the new singular term 'the property of being a dog', whose referent is the property of being a dog."

23 I'm indebted to Hartry Field for his very helpful comments on a previous draft of this chapter.

References

Boghossian, P. 1990. "The status of content," *Philosophical Review*, 99(2): 157–184.

Earman, J. J., ed. 1992. *Inference, Explanation, and Other Frustrations*. Berkeley, CA: University of California Press.

Field, H. 1972. "Tarski's theory of truth." Reprinted in Field, 2001a, pp. 3–29.

Field, H. 1978. "Mental representation." Reprinted with new postscript in Field, 2001a, pp. 30–82.

Field, H. 1986. "The deflationary conception of truth." In *Fact, Science, and Morality*, edited by G. Macdonald and C. Wright, pp. 55–117. Oxford: Blackwell.

Field, H. 1992. "Physicalism." In Earman, 1992, pp. 271–291.

Field, H. 1994a. "Deflationist views of meaning and content." Reprinted with new postscript in Field, 2001a, pp. 104–156.

Field, H. 1994b. "Disquotational truth and factually defective discourse." Reprinted in Field, 2001a, pp. 222–258.

Field, H. 2001a. *Truth and the Absence of Fact*. Oxford: Oxford University Press.

Field, H. 2001b. "Attributions of meaning and content." Reprinted in Field, 2001a, pp. 157–176.

Field, H. 2015. "Egocentric content." Unpublished manuscript.

Fodor, J. A. 1991. "You can fool some of the people all of the time, everything else being equal; hedged laws and psychological explanations." *Mind* 100(397): 19–34.

Frege, G. 1892. "On sense and reference." *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50.

Frege, G. 1918. "The thought." *Beiträge zur Philosophie des deutschen Idealismus*, 58–77.

Goldman, A. 1989. "Interpretation psychologized," *Mind & Language*, 4(3): 161–185.

Gordon, R. 1986. "Folk psychology as simulation," *Mind & Language*, 1(2): 158–171.

Grover, D., J. Camp, and N. Belnap. 1975. "A prosentential theory of truth." *Philosophical Studies*, 27(2): 73–125.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Horwich, P. 1998. *Truth*, 2nd edn. Oxford: Oxford University Press.

Jackson, F., G. Oppy, and M. Smith. 1994. "Minimalism and truth-aptness." *Mind*, 103(411): 287–302.

Kim, J. 1988. "Explanatory realism, causal realism, and explanatory exclusion." *Midwest Studies in Philosophy*, 12(1): 225–239.

Leeds, S. 1978. "Theories of reference and truth." *Erkenntnis*, 13(1): 111–129.

Ostertag, G., ed. 2016. *Meanings and Other Things: Essays on Stephen Schiffer*. Oxford: Oxford University Press.

Parret, H., and J. Bouveresse, eds. 1981. *Meaning and Understanding*. Berlin: Walter de Gruyter.

Quine, W. V. O. 1970. *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.

Quine, W. V. O. 1992. *The Pursuit of Truth*, rev. edn. Cambridge, MA: Harvard University Press.

Ramsey, F. 1927. "Facts and propositions," *Proceedings of the Aristotelian Society*, suppl. vol. 7.

Schiffer, S. 1981. "Truth and the theory of content." In Parret and Bouveresse, 1981, pp. 204–222.

Schiffer, S. 1991. "Ceteris paribus laws." *Mind*, 100(397): 1–17.

Schiffer, S. 2003. *The Things We Mean*. Oxford: Oxford University Press.

Schiffer, S. 2016. "Pleonastic propositions in moral discourse: response to Michael Smith." In Ostertag, 2016.

Smith, M. 2016. "Schiffer's unhappy face solution to a puzzle about moral judgement." In Ostertag, 2016.

Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

PART I

Language, Truth, and Reality

20

Realism and its Oppositions

BOB HALE

In many branches of philosophy, dealing with very different areas of our thought and talk, there occur disputes centered on the tenability of positions described as 'realist.' In the philosophy of science, realism stands opposed to various forms of instrumentalism; mathematical realists, often known as Platonists, are opposed in one way by nominalists, in another by constructivists; moral realists contend with subjectivist tendencies, such as expressivism and projectivism, as well as with error theories; in the theory of meaning itself, realism is under attack from positions which hold that meaning must be explained in terms which preserve an essential link between what we mean and evidence, as well as from meaning-skeptical arguments advanced by Quine, Kripke, and others (see Chapter 26, INDETERMINACY OF TRANSLATION; Chapter 23, ANALYTICITY; Chapter 31, MODALITY, §2; Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING; and Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM). It is scarcely obvious that there is some single type of issue at stake in these disputes (henceforth R/AR disputes), or that there is at least some significant continuity between them. The very diversity of the positions set against realism in these different areas might of itself be thought to point towards the opposite conclusion: that realism amounts to different things in the different cases, so that any attempt at general discussion is doomed to failure. It is not obvious, either, that the various disputes have anything much to do with the philosophy of language, or that there is any reason to expect arguments in the philosophy of language to play a significant part in their resolution.

Against these dampening thoughts may be set – besides the feeling that it is unlikely to be sheer coincidence that the same label is applied to completely disparate positions with no significant similarities whatever – at least two reasons why philosophers of language may properly take an interest in general questions about realism and the forms which opposition to it may assume. First, and most obviously, there is an R/AR dispute (or disputes) within the philosophy of language itself, centered on the tenability of realist theories of meaning. At the very least, it might be expected that scrutiny of R/AR disputes in other areas may

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

illuminate the issues here, if only through contrasts rather than parallels. But second, and more importantly, the notion that debates about other realisms – in science, mathematics, or other areas – may proceed unaffected by arguments in the philosophy of language overlooks the possibility that a successful anti-realist argument in the theory of meaning may ramify into other disputed areas (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §3).

We begin (§1) with an examination of Michael Dummett's influential treatment of these issues, which couples an attempt to identify a common form exemplified by a large, if not exhaustive, range of R/AR disputes with important arguments against a realist position about meaning which – if they are sound, and Dummett's diagnosis of what is at stake in those disputes is correct – promise to resolve the issue in the anti-realist's favor, not only in the theory of meaning itself, but across the board.[1] We then (§2) survey the principal negative arguments, advanced by Dummett and others, for semantic anti-realism. In §3, we turn to the wider question of the bearing of these arguments on R/AR disputes more generally, and review doubts about the adequacy of Dummett's general conception of their common form. Other ways in which the anti-realist case may be prosecuted are reviewed in §4: classical reductionist positions; error theories; expressivist/projectivist options and quasi-realism; and we conclude (§5) with a brief examination of the new perspective on R/AR disputes advocated in recent work by Wright.

1 Dummett's General Account of R/AR Disputes

Many traditional, and at least some currently active, R/AR disputes appear primarily to concern the existence of entities of some sort – objects of some general type, or perhaps entities which, if there are such, should be taken as belonging to some other category. Medieval realists and their nominalist adversaries, for example, were disagreed over the existence of universals – abstract entities conceived as objective worldly correlates of general terms like 'red' and 'honest' and denoted by corresponding abstract nouns like 'redness' and 'honesty.' The cardinal negative thesis of many modern nominalists has likewise been the denial that there exist any abstract entities – by which they chiefly understood properties or attributes, as opposed to the particular concrete entities they characterize, together with sets or classes. One kind of realism or platonism about mathematics is distinguished by its acceptance of numbers and sets as genuine objects, lying outside space and time but nonetheless existing independently of our thought. At least part of what is in dispute between scientific realists and their opponents is whether a satisfactory account of theoretical science requires us to see it as describing the properties of unobservable or theoretical entities such as particles, forces, and fields. Modal realists of one sort insist that there are possible worlds, distinct from but no less real than the actual world. (See Chapter 31, MODALITY, §3.)

Dummett's conception of R/AR disputes stands in sharp contrast with the model suggested by such examples. Issues between realists and their opponents are, he contends, usually best characterized *not* as disputes about the *existence of entities* of some problematic sort, but in terms of a certain *class of statements* – those distinctive of the area of thought and talk in question – which he usually labels the 'disputed' class. Further, the disagreement is not – or not primarily – over whether statements of the disputed class are true, since the anti-realist will agree that in many cases they are so; it concerns, rather, the nature or

character of the notion of truth which may be applied to them. This last point merits both emphasis and comment. A preference for formulating R/AR disputes in terms of problematic statements rather than problematic entities need, by itself, involve no significant break with the idea that those disputes centrally concern the existence of entities of certain kinds. It need not do so, because the preference might be grounded in the plausible view that general ontological questions (Do there exist so-and-sos?) reduce to, or are at least best approached as, questions partly about the logical form of some appropriate range of statements and partly about their truth-values. Thus one question at issue between mathematical Platonists on the one side and, on the other, nominalists and others is whether numbers, sets, and so on exist. Precisely because we are obviously not concerned with entities which might conceivably be objects of ostension or of any sort of perceptual encounter, or which might announce their presence indirectly through their effects, it is difficult to see how the question of their existence can be non-prejudicially approached, save by equating it with a question about truth and logical form: Are there true statements whose proper analysis discloses expressions purporting reference to numbers? General endorsement of this approach to questions of ontology is tantamount to acceptance of Frege's celebrated 'Context Principle' which, construed as a principle about reference, warns against asking after the reference of sub-sentential expressions outside the context of complete sentences (Frege, 1884, p. x and §62; Dummett, 1973a, pp. 192–196, 494–500; 1982, p. 239; 1991, chs 16 and 17; Wright, 1983, §§2, 3, 5, 8; Hale, 1987, pp. 10–14, 152–162, 228–230). Dummett is sympathetic to it. But his insistence upon treating R/AR disputes as centered on a class of statements is prelude to a quite different claim about their character. He writes:

> Realism I characterise as the belief that statements of the disputed class possess an objective truth-value, independently of our means of knowing it: they are true or false in virtue of a reality existing independently of us. The anti-realist opposes to this the view that statements of the disputed class are to be understood only by reference to the sort of thing which we count as evidence for a statement of that class … The dispute thus concerns the notion of truth appropriate for statements of the disputed class; and this means that it is a dispute concerning the kind of *meaning* which these statements have. (Dummett, 1963, p. 146)

As Dummett goes on to make clear, he thinks that the notion of a statement's having an 'objective truth-value, independently of our means of knowing it … in virtue of a reality existing independently of us' is to be understood in a very strong sense. The realist is to be understood as holding not merely that a statement may be true or false without our actually knowing its truth-value, nor even that a statement may be true or false even though we are in fact or in practice unable to tell which, but that there can be a much more radical dislocation of truth-value and our capacity for its recognition – a statement may possess a determinate truth-value without its being possible, even in principle, for us to come to know it (Dummett, 1963, p. 146; 1969, p. 358; 1973b, p. 224; 1982, p. 230). It is for this reason that realism, as Dummett conceives it, amounts to – or at least crucially involves – a thesis about meaning: to adopt a realist view of any area of thought and talk is to conceive of its distinctive statements as endowed with meaning through being associated with evidentially unconstrained truth-conditions, that is, conditions whose satisfaction bears no essential connection, however attenuated, with the possibility of its being recognized by us.

Although the foregoing characterization may be taken as definitive, Dummett very frequently depicts the issue between realists and their opponents in other, ostensibly quite

different terms, as concerning the principle of bivalence, according to which every statement is either true or false.[2] It is clear that in taking endorsement of unrestricted bivalence as 'a touchstone for a realistic interpretation of the statements of some given class,' Dummett intends no departure from his official characterization. The relations between the two are, however, by no means straightforward. It is, certainly, very plausible to regard unqualified endorsement of bivalence as *sufficient* for realism. For it is a plain fact that our language affords the means of framing various kinds of statement which are not effectively decidable – that is, statements for which there exists no procedure guaranteed to issue, after finitely many steps, in a correct verdict on their truth-values. To insist that such statements are, nevertheless, determinately either true or false would, it seems, require thinking of them as capable of being true, or false, in the absence of evidence either way, and thus as possessed of potentially evidence-transcendent truth-conditions. But realism does not obviously entail a commitment to unrestricted bivalence. It seems that one might decline to endorse bivalence for reasons which appear quite consistent with holding that certain statements may have their truth-values undetectably, say because one took failure of reference on the part of ingredient singular terms to deprive statements of truth-value.[3] A further complication concerns vagueness, which is commonly – though not invariably – taken to cause certain statements to lack determinate truth-value (see Chapter 28, SORITES). These considerations indicate that refusal to endorse bivalence may or may not signal adoption of an anti-realist view, depending upon the specific reasons for that refusal. If realism does involve a commitment to bivalence, it would seem that it can be at most a conditional one, to the effect that any statement is true or false whose ingredient terms are not subject to vagueness or reference-failure. Whether and how this qualified claim can be established, and, in particular, how it might be shown that vagueness and reference-failure are the only grounds on which a realist may properly refuse to endorse bivalence, are hard questions to which, so far as I know, we still want answers. Here they must be left open.

There are, as we have observed, many different areas in which what seems aptly described as a realist position may be defended or opposed. There is no clear presumption that one must be committed to realism across the board, if one seeks to uphold a realist position in any quarter of it. On the contrary, it appears that realism in one area might consist perfectly well with opposition to it in another – that one might, for instance, defend a realist view about theoretical science whilst rejecting realism about ethics, or values generally, or, even more selectively, combine a realist attitude towards some parts of scientific theory (such as classical physics) with anti-realism about other parts (such as quantum mechanics). Certainly there appears little prospect of a quite general argument enforcing adoption of a globally realist stance. A considerable part of the interest and importance of Dummett's configuration of R/AR disputes undoubtedly lies in the fact that it opens up the possibility – which might otherwise appear no less remote – of a quite general argument of the opposing tendency, enforcing global anti-realism across all the disputed areas. For if Dummett is right, realism everywhere depends upon the viability of a realist conception of meaning in terms of potentially evidence-transcendent truth-conditions (hereafter, 'realist truth-conditions' for brevity). Thus any argument against semantic realism as such is potentially quite generally destructive of realist options. There are, accordingly, two main questions requiring attention: (1) Are there compelling arguments – perhaps ones advanced by Dummett himself – against a realist conception of meaning? (2) Has Dummett provided an adequate general characterization of R/AR disputes? In the 50 or so years since Dummett's

earliest publications that bear on them, both questions have generated a very considerable amount of critical discussion, of which only the briefest overview can be given here.

2 Arguments against Semantic Realism

Dummett himself advances two main arguments against the idea that our understanding of disputed statements could consist in our associating them with realist truth-conditions, one focused on the difficulty of seeing how we could *acquire* such an understanding, and the other on the difficulty of seeing how we could *manifest* it. As will quickly become apparent, neither argument purports to be conclusive: each is, rather, to be seen as presenting the realist with a challenge which she appears unable to meet.

According to the *Acquisition Challenge*, our training in the use of language consists in our being taught to accept statements as true in circumstances of such-and-such a sort, and to reject them as false in circumstances of other sorts. This training *necessarily* proceeds in terms of states of affairs which we can *recognize* as obtaining. But how, in that case, are we supposed to come by the conception of evidence-transcendent truth-conditions which the realist postulates? How are we to come to know what it is for a statement of that kind to be true, or false, in virtue of the obtaining of some state of affairs which obtains *undetectably*? The challenge is to explain how we come to assign to statements truth-conditions involving states of affairs which, by their very nature, *can have played no part* in the process by which the meanings of those statements are learned or communicated. If it is conceded that there can indeed be no *ostensive* training that enables us to form such a conception, but suggested that we can nevertheless acquire it through *verbal explanation*, the counter may be given that this merely postpones the problem, since presumably no verbal explanation can be adequate that does not itself employ sentences already understood as having evidence-transcendent truth-conditions – but in that case, how is the proposed explanation to get off the ground?[4]

The *Manifestation Challenge* runs thus: If the meaning of a statement consists in its having certain (possibly evidence-transcendent) truth-conditions, then understanding it (knowing its meaning) is possessing knowledge of such. But knowledge of a statement's meaning cannot, in general, consist in the ability to provide an informative statement, in other words, of what it means (and obviously it can't consist in the ability to state *un*informatively what it means, just by disquoting it). We may concentrate on the case where knowledge of meaning does not consist in the capacity to give a verbal explanation of meaning, since no such explanation can introduce the possibility of evidence-transcendence. When knowledge of meaning is not verbalizable but implicit knowledge, it must be knowledge of how to use the sentence, and must therefore consist in the speaker's possession of certain practical abilities. But now, by just what practical abilities is an alleged grasp of evidence-transcendent truth-conditions supposed to be manifested? In the case of effectively decidable statements, or of statements which, whenever they are true, are recognizably so, a speaker's implicit knowledge can be identified with his capacity to discriminate between circumstances in which the statement is true and those in which it is not. But it clearly cannot do so in the case of any statement possessed of evidence-transcendent truth-conditions – in this case, there is nothing a speaker can do which fully manifests his supposed grasp of those conditions. Realism thus clashes head-on with the Wittgensteinian equation of meaning with use and of understanding with capacity for correct use.[5]

Attempts to Answer the Acquisition Challenge

Truth-Value Links

Among the types of statement that are problematic, in view of the anti-realist challenge, are statements about the past and about other minds. The realist conception has it that such statements can be determinately true or false in virtue of past states of affairs, or states of mind of others, to which we have no direct access, and for which adequate evidence may be quite simply unavailable. And the challenge is then to explain how we come by this conception. One suggestion is that the truth-values of statements of these problematic kinds are systematically connected with those of statements lying outside the anti-realistically problematic class – in these cases, present-tensed statements and first-person psychological statements. Thus there is a systematic link between the truth-value of a past-tensed statement made at one time, say now, and various corresponding present-tensed statements which were, or could have been, made at earlier times; for example:

> The statement: 'One million years ago, this place *was* covered with ice' is true now if and only if the statement 'This spot *is* covered with ice,' made a million years ago, was (or at least would have been) true.

The thought, then, is that understanding this truth-value link is an uncontroversial component in our mastery of tensed discourse. But present-tensed statements are not, as such, anti-realistically problematic, since they relate to conditions which obtain (or don't, as may be) detectably or recognizably. By our grasp of these two things, it is claimed, we can come to understand what it is for past-tensed statements to be true in virtue of states of affairs which are no longer accessible to us.[6]

This response fairly obviously fails to provide a *general* answer to the acquisition challenge, since no such maneuver appears feasible in the case of other types of problematic statement, such as unrestricted quantifications over an infinite, or otherwise unsurveyable, totality of objects, such as the natural numbers. Of course, '∀nPn' is true iff all its instances are true. But this is clearly no advance, since whilst the truth-value of each 'Pn' may be unproblematically recognizable, if 'P' is a decidable arithmetic predicate, we enjoy no unproblematic access to the fact, if it is one, that *all* of them are true.[7] But even in cases where the truth-value link gambit appears available, it does not really work. The trouble is that present-tensed statements have unproblematic (detectable) truth-conditions *only in the context of present use*. But the link only helps if we understand what it means to say, for example, 'This spot is covered with ice' *was* true; that is, what is *ceteris paribus* unproblematic is what it is for a present-tensed statement to be *true now*, but what we need, to move from right to left across the truth-value link to knowledge of what it is for a past-tensed statement to be now, but undetectably, true, is understanding of what it is for a present-tensed statement *to have been true* – and this is no less problematic than what we are seeking to explain.[8]

Partial Accessibility

We can distinguish between *chronically e-transcendent* statements – such as 'Everything in the universe has doubled in size' and 'The entire universe sprang into existence just five minutes ago, replete with traces of a long and complex past, etc.' – which by their very nature could in no possible circumstances be recognized as true, and statements which,

though not *guaranteed* to be so, are, *in favorable cases*, *detectably* true. Realists may concede that there is no hope of defending their distinctive conception of truth for the former, though claiming that this is no loss, since they are beyond the pale anyway; but they may insist that matters stand otherwise with the latter. Here, they may claim, if a statement of this sort is undetectably true, it is at most *contingent* that it is so. Statements of the same kind are, on occasion, recognizably true: that is, we sometimes have access to states of affairs of the kind which confer truth on them. And this, they may claim, is enough – enough to equip us with a conception of what it is for such statements to be true but undetectably so – this is just for there to obtain a state of affairs of the same kind as we have recognized to obtain in other cases. So it is, McDowell claims, with statements about the past and about the psychological states of others. Although we don't always, or even usually, have direct non-inferential access to past states of affairs, we do sometimes, through memory; and we can on occasion simply and literally observe that another is in pain or violent grief – we may see pain or grief in their face and actions, which express or manifest their state.[9]

Like the preceding response, this is of limited application at best. It is doubtfully available in the case of statements about the remote past, beyond the reach of living memory. Further, no response of this sort seems available for spatially or temporally unrestricted contingent generalizations (whether lawlike or accidental), or for quantifications through an infinite domain – in neither case does there appear to be any purchase for the idea of our being sometimes graced with direct access to an appropriate truth-conferring state of affairs. Clearly, too, the idea of occasional direct access to others' psychological states may be challenged. But there is a quite general difficulty with the partial access gambit, even in what might seem favorable cases.

First, and obviously, we should distinguish between the (problematic) case of a statement's being *undetectably* true and the (unproblematic) case of a statement's being true, though not, as it happens, known to be so, simply because we haven't taken steps we could have taken to ascertain its truth-value. We can, plausibly, understand what it is for a statement in the latter case to be true, in terms of there obtaining a state of affairs of the same kind as we have verified to obtain in the case of other statements of that type. But this is not to the point – for it is another, and much stronger, claim that we can come by the notion of undetectable truth by this route.

Second, with this out of the way, we can see that the crucial, but contentious, claim is that statements in respect of which we do *not* enjoy direct access to any truth-conferring state of affairs are *of the same kind* as other statements, such as those about the past, for which we do. Once it is allowed that they *are* of the same kind, it may seem an easy step to the realist's desired conclusion, that we can conceive of the former as true in the same way as the latter, for all that the former are, as it happens, undetectably true. Now they *are* of the same kind in one sense, for they are all statements *about the past*. But this, the anti-realist may protest, is not the point. In another sense, they are *not*: for the former are (allegedly) undetectably true, if true, whereas the latter are, *ex hypothesi*, detectably so. The realist simply *assumes*, but does nothing to show, that this difference *makes no difference*. But that it does make a difference is precisely the content of the acquisition challenge. So the question is begged, not answered.[10]

Enhanced Recognitional Capacities

The idea that underpins the preceding response, that undetectability of truth-value commonly derives from contingencies of our circumstances or contingent limitations upon our recognitional capacities, is sound enough. There may be some temptation to think it can be

exploited to the realist's advantage in a somewhat different way. It may be suggested that we can attain a conception of what it is for statements to be true – though *undetectably* so as far as *we* are concerned – by conceiving of creatures with suitably extended powers of recognition, for whom the obtaining of the relevant truth-conferring states of affairs would *not* be undetectable. Thus far, the moderate anti-realist need have no objection – indeed, unless he is able to appeal to conceivable finite extensions of our powers of recognition, computation, and so on, it is hard to see how he might resist the slide into an implausibly extreme version of anti-realism (strict finitism, in the case of mathematics), according to which the only meaningful statements are those which we can actually verify.[11] But if the suggestion is to serve the realist's ends, it must go beyond envisaging relatively uncontroversial, finite extensions of our detective abilities, to encompass conceiving, for example, of creatures capable of surveying infinite totalities, or 'directly seeing' into the past and future, or into the minds of others. It is not, however, at all clear that we can conceive any such thing in relevant detail – it is one thing to appeal to the idea of creatures whose recognitional capacities *finitely* extend powers we *actually* possess, and quite another to claim that we can conceive of creatures endowed with capacities for which we have *no* actual model, or which constitute *infinite* extensions of our capacities. It may further be objected that even if we could imagine a use for certain sentences, by beings with powers greatly exceeding our own, this does not automatically put us in a position to use those sentences in that way ourselves. To suppose that we could is to suppose that we could employ those sentences with the intention of conforming to standards of correct use, even though we are ourselves entirely unable to tell whether our use accords with them or not. This seems to run afoul of considerations concerning the *normativity* of meaning. To attach a certain meaning to a statement is, in effect, to divide states of affairs into two classes: those in which it is correct to assert the statement, and those in which it is incorrect. To employ the statement with that meaning is, prescinding from complications about insincerity, to use it with the intention of asserting it only in states of affairs of the first kind. But it appears quite generally to make no sense to suppose that an agent intends to φ if there is in principle no means available, however indirect, of telling whether or not he has succeeded in φ-ing. So in particular, nothing can be made of the suggestion that we intend, in using certain statements, to conform to standards of correctness in use, conformity with which essentially eludes detection *by us*.[12]

Compositionality

A better response might begin by pointing out that the opening claim of the acquisition argument – that our training in the use of language consists in our being taught to accept statements as true in circumstances of such-and-such a sort, and to reject them as false in circumstances of other sorts – is liable to deflect attention from the crucial point that our understanding of most sentences comes not through any directly forged link between them and recognizable states of affairs in which their assertion is justified, but is the product of prior understanding of their ingredient expressions together with their mode of construction. It may then seem that a very simple response is available: we come by a grasp of realist truth-conditions by coming to understand sentences having those truth-conditions, and we come to understand such sentences in just the way in which we come to understand the vast majority of sentences in our language, by understanding their ingredient words and semantically significant syntax.

The anti-realist must indeed accept that our understanding of sentences of the various kinds central to the dispute – statements about remote regions of space or time, quantifications

through infinite or unsurveyably large totalities, counterfactual conditionals, and so on – is, in general, acquired along compositional lines. But he will likely object that the further claim – that that understanding involves the association with those sentences of realist truth-conditions, rather than, say, conditions of justified assertion – is entirely gratuitous. The realist may, and should, concede that the proposed response does not *prove* that we understand the problematic sentences as having realist truth-conditions. But she can point out that it was not intended to do so; the aim was rather to explain, on the *assumption* that our understanding has that character, how we may have acquired it. In the absence of an argument showing that a grasp of realist truth-conditions cannot emerge, via composition, from agreed unproblematic starting points, or an independent argument – perhaps one based on manifestation – for the bankruptcy of the distinctive realist conception of truth-conditions (of which it would be a corollary that we cannot have acquired it), it seems that this is enough to neutralize the acquisition challenge.

Attempts to Answer the Manifestation Challenge

Explanatory Ascription

Several critics claim that the manifestation challenge may be met, or deflected, by observing that the ascription of knowledge of realist truth-conditions may form part of a successful theoretical explanation of speakers' behavior. A closely related suggestion is that knowledge of meaning is manifestable in a capacity to interpret the speech-behavior of others, where this involves, centrally, the correct ascription of beliefs – which may be realist beliefs – which figure, in combination with suitable desires, in explanations of speakers' behavior.[13] But there is an obvious difficulty with this kind of response. Evidently there is no reason why an anti-realist should not go in for interpretations of linguistic behavior, or explanations of behavior in general, in terms of beliefs. It is therefore essential to show, if the proposed reply is to make headway, that such interpretations and explanations must sometimes proceed in terms of specifically realist beliefs.[14] That is, it needs to be indicated what specific aspects of behavior, or the capacities to which they bear witness, call for explanation in terms of the hypothesis that the subjects of ascription hold beliefs, the content of which demands characterization in terms of realist truth-conditions rather than conditions of warranted assertibility, say; otherwise, the ascription of realist beliefs will merely incorporate so much theoretical slack.[15]

Inferential Practice

It may be suggested that a realist understanding of certain statements may be manifested by our employment of distinctively classical principles of reasoning in our willingness to reason by (unrestricted) use of the Law of Excluded Middle, or by Double Negation Elimination, or other patterns of inference rejected by the intuitionists and unjustifiable save on the assumption that the statements are apt for evidence-transcendent truth.

A difficulty with this is that it involves treating our actual inferential practice as sacrosanct. As against this, it may be said that the justification for employing certain principles of inference should be given in terms of the kind of meaning we have conferred upon the statements involved, so that the appeal to inferential practice gets things the wrong way round. Why should we take our unrestricted use of classical logic as showing that we have conferred realist meanings for the sentences involved, rather than as revealing that we have – by a kind of uncritical inertia – illegitimately projected patterns of reasoning that are

correct within a restricted domain (say, of decidable statements) to cases which lie outside it? Even if it is granted that a propensity to reason classically betokens a *commitment* to assigning realist truth-conditions to problematic statements, it does not seem that this could suffice to make clear what specific truth-conditions have, putatively, been assigned to them.[16]

Other Modes of Manifestation?

In case of decidable statements, a speaker may manifest a grasp of truth-conditions in a quite straightforward way, by implementing the appropriate procedure, leading to her recognition of the statement as true, or false, as may be. Where a statement is not effectively decidable, and is in fact true but not recognizably so, a speaker cannot, obviously, manifest a grasp of its truth-condition by determining its truth-value. But the thought is tempting that she may nevertheless demonstrate an appropriate understanding in other ways. Thus Strawson (1977, p. 16) proposes that whilst it is, of course, true – and a truism – that 'grasp of the sense of a sentence cannot be displayed in *response* to unrecognizable conditions,' it will be

> enough for the truth-theorist that the grasp of the sense of a sentence can be displayed in response to *recognizable* conditions – of various sorts: there are those which conclusively establish the truth or falsity of the sentence; … those which (given our general theory of the world) constitute evidence, more or less good, for or against the truth of the sentence; … even those which point to the unavoidable absence of evidence either way.

His thought is that there are various responses to recognizable states of affairs which can be regarded as manifesting a grasp of the sense of a sentence, in addition to recognizing that the condition for its truth definitely does, or definitely does not, obtain.

So there surely are, but this does not seem enough. Here it is crucial to remember that the truth-theorist to whose defense Strawson is (or ought to be) contributing is a realist, who holds that grasp of the sense of a sentence consists, in the case where the sentence is not effectively decidable, in knowledge of its possibly e-transcendent truth-condition. The responses Strawson mentions, however, are entirely consistent with the anti-realist view that, in such cases, understanding the sentence consists in knowing the conditions for its warranted assertion. That is, such responses do not distinctively display grasp of *realist* truth-conditions for the sentence.[17]

Manifestation and Manifestees

Manifestation is a relational matter. A chess master may be able to manifest her skills to others with a reasonable knowledge of the game; but she cannot be expected to display all or even any aspects of her virtuosity to those unfamiliar with it. Simon Blackburn claims that the manifestation argument only appears compelling because it tacitly restricts the manifestees to cognitively impoverished creatures, capable only of *observation*. Thus, picking up on Dummett's remark that 'an individual can communicate only what he can be observed to communicate' (Dummett, 1973b, p. 217), Blackburn takes this to suggest that the manifestee is to be

> one who is capable of making observations, but no more. But let us suppose that some things lie outside observation: the past, or other people's sensations, or sub-atomic particles. Then it is clearly not a sensible requirement that a man should manifest his understanding of these

> things to someone who is capable of only making observations.… Such limited observers make poor audiences…. the very word 'manifest' reveals the doubtful nature of this requirement. Like 'display' and 'reveal,' it has largely visual overtones: I cannot display or make visible the past events I talk about, the future ones, my own pains and thoughts, let alone electrons or numbers. (Blackburn, 1984, pp. 65–66; 1989)

But this complaint, it seems to me, misrepresents Dummett's argument. That an anti-realist imposes no restriction to states of affairs which can be observed to obtain, when they do, should be clear from the mathematical case which forms the departure point for his work. Dummett raises no problem about finitary mathematical statements, although their truth, if they are true, plainly does *not* consist in the obtaining of some *observable* state of affairs. The important distinction is not between the observable and the unobservable, but between effectively decidable statements and others. In the case of any statement whose truth or falsity is an effectively decidable matter, we can equate implicit knowledge of the statement's truth-condition with a capacity to decide it. The case where the statement concerns some literally observable state of affairs is just a special case of this – here there is a particularly simple decision procedure: just look and see (sniff and smell, and so on). It is with non-effectively decidable (non-ED) statements that the problem arises: their being non-ED does not, of course, mean that there cannot be circumstances in which we are able to tell that the condition for their truth is fulfilled or not. Thus to take a famous example, if Goldbach's Conjecture that every even number is the sum of two primes is false, we might one day be confronted with a counter-example; and there is no reason to suppose that, if it is true, it is nevertheless insusceptible of proof – so we could conceivably be in position to recognize it as true. The point is that we have no effective way of bringing about a situation of either kind. It is for this reason that we cannot equate a knowledge of the statement's truth-condition (as the realist conceives it) with a capacity to recognize a proof of it, or a counter-example to it, should we be lucky enough to find ourselves confronted with such. For then the statement would be true if and only there is a proof of it, and false if and only if there is a disproof. But there is no guarantee that a situation of either kind will ever obtain, so the realist could not be justified in taking it that the statement is true or false all the same (cf. Dummett, 1976, pp. 81–82).

In particular, the suggestion that talk of what can't be made visible – such as numbers – is threatened by the manifestation requirement quite misconstrues its intended force. What has to be capable of manifestation is our supposed knowledge or understanding, not the objects we talk about. If Blackburn's reading of Dummett were right, he ought to find *all* statements about numbers problematic. But of course, he does not: there is no special difficulty in saying what (implicit) knowledge of the truth-condition for example '937 is prime' consists in – it consists in mastery of a procedure for deciding the statement.

Blackburn's key claim – that the argument relies on some unduly restrictive assumption about the capacities of suitable manifestees – is thus indefensible as a reading of Dummett's texts. It might nonetheless be claimed that the argument has plausibility only under some such restriction. But is that so? Let's suppose the audience is as competent in the use of the statements in question as the speaker. That is, grant what Blackburn would reckon as a suitable audience – someone competent in number theory, say. We can still ask: What recognizable capacities of the speaker constitute his supposed knowledge of what it is for statements involving unbounded quantification over the natural numbers to be true, but undetectably so? Bringing in a suitable audience appears in itself to make no advance on

the problem – we still need an answer to the question (cf. Wright, 1993a, pp. 20–21). This discussion of the principal anti-realist arguments has been quite selective,[18] and is, it will be only too clear, inconclusive. My aim has been limited to providing little more than an introductory survey of some of the main moves made in this complex, difficult, and, in my view, still unresolved debate.

3 The Adequacy of Dummett's Characterization of R/AR Disputes

If Dummett's account of R/AR disputes in general is acceptable, the potential bearing of anti-realist arguments in the theory of meaning upon their resolution is immediate: if successful, such arguments would enforce a globally anti-realist stance. But is it acceptable? Dummett has never maintained that his characterization fits *every* dispute which might be taken to concern the tenability of position describable as realism (cf. Dummett, 1963, pp. 146–147). His claim has rather been – and continues to be – that it captures what is centrally at issue in an extensive range of such disputes. Few would deny that there are cases to which Dummett's characterization seems entirely apt. The dispute between Platonists and constructivists in the philosophy of mathematics is probably the clearest example. Platonists in this sense uphold – and various species of constructivist challenge – the legitimacy of employing in mathematical reasoning forms of inference enshrined in classical logic, such as Double Negation Elimination and a strong form of *reductio ad absurdum*, which depend for their justification upon the assumption of unrestricted bivalence, and hence upon taking mathematical statements to have realist truth-conditions. But it may be doubted whether the same is true of other R/AR disputes which have commanded interest and attention in recent and ongoing philosophical discussion.

We can distinguish three component claims in Dummett's configuration of R/AR disputes:

1. such disputes are best understood as concerned with a class of problematic *statements*, rather than with a class of problematic *entities*;
2. more specifically, what is in dispute is the *character* of the notion of *truth* which may properly be taken to have application to the statements in question;
3. more specifically still, what is primarily at issue is whether statements of the kind in question may defensibly be held to be capable of being true in a *potentially evidence-transcendent* manner.

Clearly each of these claims presupposes the correctness of – and is, in that sense at least, stronger and more contentious than – its predecessors. Our review of grounds on which the adequacy of Dummett's approach may be doubted can conveniently be organized in terms of them. We begin with some considerations relating to the first claim.

In one of his most recent papers on the present topic,[19] Dummett reiterates his conviction that R/AR disputes are best seen as concerning a class of problematic statements, rather than a class of problematic objects, and offers two reasons for it. The first is that in some cases – the examples given are disputes over the reality of the past or over the future – there is no germane class of objects for the dispute to be about. The second is that even in cases where there is some problematic type of object, as with the Platonist/constructivist dispute

about mathematics, it is not the existence of the objects as such that the dispute really concerns, but the objectivity of the statements we make about them. A philosopher who accepts a problematic class of objects – numbers or sets, for example – may yet take an anti-realist view of facts about them; while one who takes objects of that kind to be mind-dependent, perhaps because he views them as products of our intellectual activity, much as Dedekind viewed numbers as our 'free creations,' may yet be a realist, in Dummett's sense, about truths concerning them (as Dedekind appears to have been). The significant difference, Dummett claims, is between those who hold that mathematical statements, say, have determinate truth-values independently of our capacity to ascertain them, and those who deny this (Dummett, 1963, pp. 145–147; 1993a, pp. 464–465).

It is natural to protest that, unless it is intended merely as a forceful expression of Dummett's own greater interest in one kind of disagreement than another, the claim that *the* significant difference concerns one's attitude towards recognition-transcendent truth-value is tendentious. It is, of course, *a* difference, and an important one. But granting that much is perfectly consistent with acknowledgement that the disagreement between Platonists and nominalists over the existence of numbers, sets, and other kinds of abstract objects centers upon another, equally significant difference, reflecting a different aspect of R/AR disputes. More generally, it is an indisputable historical fact that some R/AR disputes are, at least in part, disputes about the existence of one or another kind of problematic entity. Obvious examples, besides the Platonist/nominalist dispute just mentioned, are the disagreements between realists and their opponents in the philosophy of science over the existence of unobservable entities postulated in advanced scientific theories, and in the philosophy of modality over the existence of possible worlds. The present point is that, even if such disputes are in part about evidence-transcendence, they involve ontological disagreements as well. If neglect of the latter aspect is indeed a consequence of Dummett's reconstruction, that is surely a serious limitation.

In fact, matters are less straightforward than these remarks suggest. Striking a more concessive note, Dummett suggests (1963, p. 147) that a dispute over the existence of certain entities might be represented in his way – that is, as concerned with a problematic class of statements – simply by taking the disputed class to consist of statements purporting reference to those entities. This suggestion accords well with the view briefly adumbrated above (Dummett, 1963, p. 273; cf. also Wright, 1993a, pp. 8–9) that ontological questions are best treated as questions about the logical form and truth-values of some appropriate range of statements. The possibility of thus reconfiguring questions about the existence of entities of some kind as questions about a certain suitably chosen range of statements is certainly enough to show that, if Dummett's approach does indeed involve an unwanted neglect of ontological issues, that is not a defect for which claim (1) is to be held responsible. But that is not, on the face of it, enough to dispose of the charge altogether. For the questions relevant to resolving the ontological issue – questions about the logical form and truth-values of certain statements – are, it seems, quite distinct from those upon which, in line with claims (2) and (3), R/AR disputes should, in Dummett's view, be concerned – questions about the character of the notion of truth applicable to statements in the disputed class. As against this, it might be claimed that what is typically at issue in a philosophical dispute about the existence of entities of some kind is not simply whether or not there are such things as numbers, say, or colors, but whether the entities in question enjoy an objective existence, independent of our thought and talk of them. But the question whether objects of some sort are objective or mind-independent – or so it may plausibly be held – is best

regarded as being whether statements purporting reference to those objects are capable of objective, mind-independent truth. Thus questions about the character of the notion of truth having application within the appropriate range of statements are, after all, central to ontological disputes, and the charge that Dummett's approach entirely neglects such disputes is therefore ill founded.[20]

This last line of defense is, I think, only partially successful. It is certainly plausible that some appropriate notion of objective, mind-independent existence is involved in (most) ontological disputes. And it is hardly less plausible, in my view, that this notion is best elucidated in terms of the idea that statements purporting reference to entities of the problematic kind are apt to be objectively true. But the claim that the notion of objective truth thus involved in ontological questions is invariably the notion of potentially evidence-transcendent truth is surely mistaken. The truth of effectively decidable statements of elementary number theory, taken at face value, suffices for the existence of indefinitely many natural numbers; but such statements, by their very nature, cannot be candidates for evidence-transcendent truth. This suggests that there has to be some other notion of objective truth, falling short of realist truth in Dummett's sense. We shall return to this point.

Turning now to claim (2), it may be felt that even if reshaping R/AR disputes as concerned with statements (rather than entities) imports no serious loss of generality, its exclusive focus upon the character of the notion of truth having application to them is a good deal less harmless: that the result is to lose contact with what is at issue in earlier disputes about realism. Dummett replies to this charge in (1993a). Traditional opposition to realism, he observes, commonly takes the form of *reductionism*. The anti-realist about a given area – schematically, the ostensible subject-matter of A-statements – maintains that there are no distinctive A-facts: rather, A-truths can be translated or paraphrased without loss or residue into B-truths, truths of some other kind which enjoy an (at least relatively) unproblematic ontology and epistemology. Thus the behaviorist denies that there are distinctively mental facts; there are just facts about overt behavior and circumstances in which it occurs, and truths about minds translate into (complex, subjunctively conditional) truths of the latter sort. According to instrumentalists and operationists – traditional scientific anti-realists – there is no special class of truths about the unobservables of scientific theory, such talk being merely convenient shorthand for talk about observables. The fatal objection to anti-realisms of this reductionist stripe, Dummett reminds us, has been that the translation programs they enjoin simply cannot be carried through; reductive behaviorism, for example, runs aground over the holistic character of discourse about beliefs, desires, and other mental states (see Chapter 15, HOLISM, and Chapter 12, TACIT KNOWLEDGE, §2), while traditional instrumentalism fails because a suitable division cannot be sustained between observation statements and theoretical ones. But the opposing realisms, he argues, have enjoyed too easy a victory. Intuitionists in mathematics accept the irreducibility of mathematical statements, but oppose realism (as embodied in standard classical mathematics, with its reliance upon unrestricted bivalence) by insisting that their content is exhausted by an account of their justification conditions. By taking this as our model, we can see that there is space for non-reductionist anti-realist positions about the mind, or about science, or in other areas, which involve no commitment to such doom-laden translation programs. The charge of irrelevance to traditional disputes over realism is thus quite misplaced: what has been done, rather, is to disclose anti-realist options in those disputes which had not been sufficiently noticed.[21]

As against that specific charge, this is a passably effective reply. But misgivings relating to claim (2) may still be felt on a somewhat different score. Let it be granted that focus on the character of the notion of truth applicable to certain statements is well adapted to plot one kind of non-reductionist opposition to realism (and set aside, *pro tem*, the question how widely available this kind of anti-realism is): it appears, nevertheless, quite ill adapted to accommodate other, equally non-reductive forms which opposition to realism may, and in significant cases does, assume. Two such directions of anti-realist thinking, both well represented in subsequent discussion, spring readily to mind, neither of which is happily construed as occupied with the character of the notion of truth having application within the discourses they concern.

There is, first, the view – paradigmatically exemplified in the emotive theory of ethical discourse embraced by some logical positivists, and foreshadowed in the writings of Hume (cf. Hume, 1739, bk. III, pt. I. §II; Ayer, 1946, introduction, pp. 20–22, and ch. VI) – that the seemingly fact-stating, descriptive utterances characteristic of a given region of discourse are not genuine assertions at all, but are rather to be understood as expressive of feelings (whether of approval or disapproval, admiration or distaste) which we project onto the natural goings-on by presenting them in assertoric or propositional style. This projectivist species of anti-realism is to be sharply distinguished from a crude subjectivism according to which ethical utterances are sincere or insincere, true or false *reports* of morally relevant feelings. In its original and purer form, at least, it maintains not that ethical utterances fail to comply with standards of objective truth-telling, but that they are not apt for evaluation as true or false at all. While it finds its most natural, and perhaps its most plausible, application in connection with morals, and evaluative discourse in general, it admits – or is often taken to admit – of extension to other areas. Hume himself may be seen as commending a projectivist treatment of causal necessity; others, following his lead, have advocated similar treatments of other areas – for example, of modality in general (see Chapter 31, MODALITY, §§3.1 and 3.4), and even of mathematics (cf. Blackburn, 1984, pp. 210–217; 1986, passim).

There is, second, a quite different – indeed, opposed – direction of theorizing which accepts the sentences of a problematic discourse as vehicles of genuine statements, aimed at truth, but denies that they can ever attain to it on the ground that reality fails to furnish objects, properties, or states of affairs of the kinds their truth demands. Versions of this species of anti-realism – error theories, as they are often called – have been advocated in relation to moral discourse by John Mackie and to mathematics by Hartry Field; eliminativist doctrines about ordinary or 'folk-'psychological discourse are perhaps also best viewed as error-theoretical.[22] In sharp contrast with more traditional, reductionist forms of anti-realism, it is held that the statements of the discourse do indeed carry the ontological commitments their surface syntax suggests – to distinctively non-natural moral properties and states of affairs, for example, or to numbers and sets – but that, precisely because there are in reality no such things, those statements are quite generally false.

Whether either of these approaches should ultimately be reckoned the best, or even a sustainable, direction anti-realist intuitions might assume is a question to which we shall return. The present point is, quite simply, that they constitute *prima facie* playable options for opponents of realism who want no truck with orthodox reductionist strategies – naturalism in ethics or a program of reinterpreting mathematics as concerned solely with nominalistically acceptable concrete entities – but they are options which can, it seems, find no place in Dummett's configuration of R/AR disputes.[23] Their availability tells, in the first instance, against claim (2). However, since neither the error-theorist's not the expressivist's

quarrel with the realist concerns the possibility of holding statements in the disputed class to be capable of evidence-transcendent truth – or subject to bivalence – these examples tell also, albeit indirectly, against claim (3).

A more direct objection to claim (3) – or to taking endorsement of bivalence as the hallmark of realism – focuses upon the phenomenon of vagueness. On one widely accepted view, vague statements are precisely ones which lack determinate truth-values. If that is so, then Dummett's characterization would seem to leave no space for any form of realism about vague discourse (cf. Wright, 1993a, p. 4). Given the very considerable extent to which vagueness pervades our language, this would constitute a serious limitation. But it is not clear that the objection is decisive. One quite radical response to the objection would be to reject the assumption on which it proceeds, that vagueness involves lack of truth-value. Such is the burden of the epistemic conception of vagueness, which has received some ingenious and determined support.[24] In this view we hesitate or are reluctant to assert, in problematic cases, that a certain colored patch is red or that a certain man is thin, not because these statements lack determinate truth-values, but because we do not know what those truth-values are. As Dummett observes, this requires us to hold that our use of vague expressions "confers on them meanings which determine precise applications for them that we ourselves do not know."[25] Finding this supposition implausible, Dummett offers a quite different answer to the objection, contending that the realist should hold that, for every vague statement, there is a range of more precise statements exactly one of which is true and the rest false, while an anti-realist is free to deny this. His thought, it seems, is that while a realist need not endorse an unrestricted principle of bivalence, he can only allow failures where they can be put down to an eliminable lack of precision on our part: if a statement lacks truth-value, that is due not to any indeterminacy in reality but to some looseness in our description.[26] Whether this conception involves no significant departure from the idea that realism is to be characterized in terms of commitment to the possibility of evidence-transcendent truth, or whether, alternatively, unrestricted bivalence may – at least as far as the difficulty over vagueness is concerned – be retained as the mark of realism by upholding an epistemic conception, are delicate questions we shall not try to resolve here.

4 Error Theories, Projectivism, and Quasi-realism

Even if an affirmative answer to either of these questions can be sustained, there are, as remarked, lines of attack apparently open to anti-realists which seem not to fit comfortably into Dummett's general characterization of R/AR disputes, since the relevant anti-realist thesis is not that the problematic statements cannot have evidence-transcendent truth-values, but that they are either not really up for assessment as true or false at all (the expressivist/projectivist option) or invariably false (error theories). It may begin to seem that it is not just Dummett's particular focus that distorts or oversimplifies, but that any attempt to identify some one kind of thing that is at issue in all R/AR disputes is unlikely to do justice to the variety of forms opposition to realism – and, correlatively, realism itself – may take. We shall consider in the next section whether the prospects for an illuminating overview are as bleak as our discussion so far suggests. First, it will be convenient and instructive to review, if only and inevitably somewhat briefly and provisionally, the error-theoretic and expressivist options, along with some more sophisticated variants of them.

In its starkest and most uncompromising form, error-theoretic anti-realism would seem to enjoin rejecting the problematic discourse outright, as resting upon presuppositions which, if its negative ontological claims are correct, are recognizably unfulfilled. The continued practice of making moral distinctions, coupled with recognition that the world fails to provide states of affairs of the kind which that practice, properly understood, demands would amount at best to bad faith. If that is a consequence we should find it hard to swallow in the moral sphere, its analogue in the mathematical case seems, if anything, even more clearly intolerable, given the apparent indispensability of mathematics to successful theorizing about the world. For this reason, much interest and importance attaches to the possibility of mitigating the apparently disastrous consequences of pure error-theoretic anti-realism by combining it with a supplementary theory which would explain how, error notwithstanding, we may be rationally justified in continuing to practice the discourse in question. Field (cf. Field, 1980; 1989) may fairly be viewed as arguing for just such a modified error theory about mathematical discourse. Mathematical statements, taken literally and at face value, are indeed, by his nominalist lights, systematically false, simply because the world fails to contain the numbers, functions, sets, and so on required for their truth.[27] But it is, in his view, a further error to suppose that this deprives them of any respectable employment. Mathematics does not need to be true to be good. What is required – and all that is required – to justify the use of mathematical theories in everyday or scientific theoretical contexts is that such theories should be *conservative*, in the sense that relying upon them in reasoning about non-mathematical matters does not enable us to reach any non-mathematical conclusions from non-mathematical premises which are not logical consequences of those non-mathematical premises alone.[28] If this idea can be made to work, it promises at least one significant advantage over more traditional forms of nominalism, by doing away with the need for the kind of reductive translation program in which – with less-than-encouraging prospects of success – they standardly engage. Whether it *can* be made to work is too large a question for adequate treatment here, but one serious-looking problem merits brief mention. This concerns the belief in conservativeness which supplants a belief in truth as the core of this anti-realist position. The question is whether Field can give a satisfactory account of its content without destabilizing his nominalism. To be conservative, a mathematical theory must be consistent. Since a theory's consistency cannot be explicated in the usual model- or proof-theoretic terms without breaking faith with nominalism, Field must take it to consist in the possibility that its axioms are collectively true. But they are in fact false in his version of nominalism. The implausible upshot would seem to be that the existence/non-existence of numbers and such like must be, in Field's view, not merely a pure contingency, but a metaphysically brute one. Any purported explanation, in nominalistically acceptable terms, of its resolution – either way – would locate non-mathematical states of affairs which would have been otherwise, had the alleged contingency been otherwise resolved, and would thus be in tension with the conservativeness of mathematical theories. The objection, it should be stressed, is not to the notion of brute contingency as such, but to the idea that the (non-)existence of numbers and sets may be properly regarded as exemplifying it.[29]

Expressivism about a discourse avoids this particular difficulty, since it denies that its utterances are genuine statements, properly assessable as true or false. But in its pure form, it runs into others. Typically, the problematic utterances will exhibit many, if not all, of the main features – with, of course, the exception, if expressivism is right, of a capacity for truth-value – of genuine assertion. Moral and modal utterances, for example, happily

tolerate embedding under negation, within disjunctions and conditionals, and in reports of propositional attitude. They are at least – however misleadingly – *said* to be true, or false. And, most importantly, they may figure – both atomically and under such embedding – as premises or conclusions of what are, to all intents and purposes, deductively valid inferences. Since Geach (1965) first drew attention to the fact, it has been a commonplace that expressive theories encounter grave difficulty in doing justice to these indisputable aspects of use, as they must if they are to be credible. Whatever attitude or sentiment is expressed by a free-standing utterance of 'Lying is wrong,' say, those same words can no longer be held to express it when they figure as antecedent to a conditional, such as 'If lying is wrong, so is getting others to lie.' On the face of it, the antecedent position needs filling with words articulating a condition which may or may not be met, and so by words apt to express not a feeling but a truth. In addition, if 'Lying is wrong' does no more than express a feeling when uttered on its own, it is difficult to see how the ostensibly valid inference from it, together with the above conditional, can be anything other than a crude equivocation.

A sophisticated development of the expressivist/projectivist approach has been forcefully advocated by Simon Blackburn. It is a central objective of his *quasi-realism* to show that acknowledging the expressive/projective basis or origin of moral discourse – to take the case for which his view is most fully worked out – need force no admission that our tendency to talk and think *as if* moral judgments are genuine assertions, having truth-conditions, is misplaced or defective. In his words, quasi-realism seeks to show "that even on anti-realist grounds, there is nothing improper, nothing 'diseased' in projected predicates … it tries to earn, on the slender basis [i.e., of projectivist assumptions], the features of moral language (or of the other commitments to which a projective theory might apply) which tempt people to realism."[30] Earning the right to present our moral or other evaluative commitments in propositional style – as Blackburn would put it – evidently requires, *inter alia*, solving Geach's problem. Blackburn's proposal is that when we assert conditionals with evaluative components, like 'If lying is wrong, so is getting others to lie,' we are expressing complex, higher-order evaluative attitudes – in this case, of approval towards combining disapproval of lying with disapproval of getting others to lie. This, he hopes, enables us to explain what is going on when we make evaluative inferences, such as the moral *modus ponens* we have taken as example. Someone who sincerely endorses the premises approves of combining disapproval of lying with disapproval of getting others to lie, and disapproves of lying. She ought, therefore, to disapprove of getting others to lie (thereby sincerely endorsing the conclusion), since if she does not she will be involved in a kind of attitudinal inconsistency – her attitudes will clash with one another, as Blackburn puts it.[31]

This fails to do justice to the problem. Someone who declines to accept the conclusion that getting others to lie is wrong from the given premises is to be convicted of *logical* incompetence, and not a merely *moral* fault (failing to have a combination of attitudes of which one approves), as on Blackburn's account of matter.[32] There is, in any case, room for doubt whether the approach can be extended to cover the full range of utterances which are, in Blackburn's view, ripe for projectivist-cum-quasi-realist treatment. It is unclear, for example, that it can deal satisfactorily with 'mixed' conditionals and other compounds involving genuinely factual components alongside evaluative ones, such as 'If Henry said that, he ought to apologize.'[33] Again, a projectivist/quasi-realist treatment of modality has it that 'It is necessary that p' functions to express our own imaginative limitations – something

like 'inability to make anything of a possible way of thinking which denies [that p]' (cf. Blackburn, 1984, p. 217) – but there seems little prospect of dealing with iterated modalities in this style.

5 Realism and Objective Truth

Let us take stock a little. We have largely been concerned to assess the adequacy of Dummett's characterization of R/AR disputes. Our discussion suggests that whilst some such disputes are indeed to be seen, as Dummett recommends, as turning upon the tenability of a conception of truth and falsity as potentially evidence-transcendent, there are others where this is not the issue. It seems, for example, that there should be space for a position appropriately describable as realist about (even) effectively decidable mathematical statements. Realists about morals, or about modality, need not, it seems, embrace evidentially unconstrained conceptions of moral or modal fact simply by virtue of their opposition to error-theoretic and expressivist/projectivist accounts of the subject-matter. It merits emphasis that what these and other examples call in question is not, as such, the aptness of Dummett's depiction of realism as 'the belief that statements of the disputed class possess an objective truth-value, independently of our means of knowing it: they are true or false in virtue of a reality existing independently of us': this remains – at least for anyone who accepts that the issues are best seen as primarily concerning a class of problematic statements, rather than a class of problematic entities – as good a schematic characterization of the position as we might hope to give. Rather, what they tell against is the particular, and particularly exacting, interpretation Dummett imposes upon its key ingredient terms – 'objective truth-value,' 'independent of our means of knowing,' and 'independent reality.' Encashing these notions in terms of potential for evidence-transcendent truth undoubtedly hits off one very strong sense in which the states of affairs a given kind of statement purportedly represents may be held to be objective and independent of our talk and thought. But might there not be other, less demanding but still substantial, conditions whose satisfaction by (true) statements about a given subject-matter would suffice for the correctness of (a form of) realism about them? An affirmative answer would invite acceptance of one broadly negative moral: that there need be no one, unique mark of realism uniform across all R/AR disputes – no one thing that is at issue in all of them. But that need not be seen as enforcing the disappointing conclusion that we are confronted with no more than a rag-bag of disparate oppositions, with nothing but a label in common. It need not do so, at least, provided we may view the relevant conditions as reflecting features of – or constraints upon the application of – the notion of truth properly deployed in the various regions of discourse over which realists and anti-realists of one or another kind may disagree. This, in broad outline, is the overall picture recommended by Crispin Wright, of which we now provide a brief review.

In contrast with expressivists, Wright argues that anti-realists about a given discourse should acknowledge, along with their realist opponents, that its distinctive utterances are genuinely assertoric and so apt for evaluation as true or false; and that, in contrast with error theorists, there is no systematic reason why its assertions should fail to be true. As Wright recognizes, this makes urgent two questions: How can the concession that the relevant utterances are truth-apt, and that many of them are indeed true, 'avoid giving

the game to the realist straight away' (unless coupled with some form of reductionist reconstruction of those statements)? And, assuming a satisfactory answer to that question, what *is* in dispute between realists and anti-realists, if both parties agree that the problematic statements are not only truth-apt, but in many cases actually true (as they stand, without benefit of some reductive analysis in terms of statements of some other discourse)? What is left for them to disagree *about*? To answer the first question, Wright circumscribes a notion of truth – *minimal* truth, as he calls it – which is neutral between realists and their opponents. To answer the second, he identifies a number of truth-related issues over which parties who agree on the minimal-truth aptness, or minimal truth, of certain statements may yet diverge – issues where what is in question can be seen as a feature whose possession by those statements would constitute their being *substantially* true in one or another of the ways suggested by the familiar realist idioms of objective or mind-independent truth, or truth in virtue of some sort of correspondence or fit with external reality.

Minimalism

If minimal truth is to serve Wright's purposes, it must be, as he puts it, a 'metaphysically lightweight' notion, unencumbered by any of the features which import realist commitments. But it cannot, he argues, be the metaphysically *weightless* notion which a distinguished tradition initiated by Ramsey and including Wittgenstein, Ayer, and most recently Horwich (Ramsey, 1927: Wittgenstein, 1922; 1953; Ayer, 1946, pp. 78–90; Horwich, 1990) has taken truth to be. According to this conception – *deflationism* – the truth predicate stands for no real property, but is no more than a mere device of disquotation. That is, the effect of applying the truth predicate to a name of a sentence – say, one formed by enclosing the sentence in quotation marks – is to produce a sentence which says no more and no less than can be said by asserting that very sentence on its own without surrounding quotation marks. Thus on this view, the whole meaning of the truth predicate is exhausted by the Disquotation Schema:

(DS) "P" is true if and only if P

Wright argues that deflationism is unstable, because, when coupled with the seemingly undeniable assumption:

(Neg) Every statement, P, has a negation, not-P

it entails inconsistent claims concerning the relations between truth and warranted assertibility. Both are clearly norms of assertoric discourse, in the sense that in making assertions, we aim at truth, and likewise aim to make assertions we are warranted in making. Deflationism is committed, via its endorsement of (DS), to accepting the normativity of the truth predicate, in the sense that reason to think a sentence true is reason to assert or accept it. Indeed, (DS) entails that 'true' and 'warrantedly assertible' *coincide* in *normative force,* in the sense that reason to think a statement true is reason to think it warrantedly assertible, and conversely. But deflationism is also committed, by taking the content of the truth predicate to be exhausted by (DS), to holding that the truth predicate is simply a device of assertoric endorsement and hence to denying that truth is a norm of assertion distinct from being warrantedly assertible. However, it follows from (DS), together with (Neg), that truth

and warranted assertibility must be distinct norms of assertion, since the predicates expressing them can diverge in extension. For it follows from (DS) together with (Neg) that:

(1) "It is not, the case that P" is true if and only if it is not the case that P

And contraposing on (DS) itself, we have:

(2) It is not the case that P if and only if it is not the case that "P" is true

whence, by the transitivity of the biconditional:

(3) "It is not the case that P" is true if and only if it is not the case that "P" is true

But the result of replacing 'is true' by 'is warrantedly assertible' in (3) is clearly incorrect. Provided that a state of information is possible which is neutral with respect to P – that is, which fails to warrant P or its negation – the resulting biconditional fails right-to-left: it may be that neither "P" nor "It is not the case that P" is warrantedly assertible.[34]

Wright contends that the combination of coincidence in normative force with, but potential divergence in extension from warranted assertibility is not only a necessary condition for something to be a genuine truth predicate, but that it is also sufficient. This is (one way to formulate) what he understands by minimalism about truth. It is worth emphasizing that minimalism about truth is, by itself, perfectly consistent with expressivist or error-theoretic anti-realism about a discourse, since it remains, so far, open to the expressivist to deny that the discourse's utterances are genuinely assertoric, and open to the error theorist to grant their assertoric status but still deny that any of them qualify for even minimal truth. Wright does in fact reject both positions. His rejection of expressivism rests upon the further claim that we should adopt a similarly minimal conception of assertion, according to which it suffices for the sentences of a discourse to be assertoric that they exhibit the appropriate syntactical features (embedding under negation, as antecedents of conditionals, in contexts of propositional attitude reports, and so on) and are subjected to 'communally acknowledged standards of proper use' or what he often calls 'discipline.' His rejection of error theories is more qualified, and the reasons for it too subtle and complex for discussion here (see Wright, 1992, pp. 86–87).

There is one final point that needs to be made about Wright's minimalism if his positive suggestions about what is or should be at issue in R/AR disputes, to which we shall shortly turn, are not to be misunderstood. This is that the minimalist conception is not put forward as an *analysis* of truth, in direct competition with traditional 'theories of truth' like the correspondence and coherence theories (see Chapter 21, THEORIES OF TRUTH). The claim is, rather, that *any* predicate which both satisfies the Disquotation Schema and exhibits certain features – ultimately those enshrined in or derivable from the 'platitudes,' as Wright describes them, that to assert a statement is to present it as true, and that any truth-apt content has a significant negation which is likewise truth-apt – should thereby qualify as *a* truth predicate. It is therefore consistent with acknowledging that there is, or even must be, more to say about the content of any predicate endowed with those features. It is, further, consistent with a certain kind of pluralism about truth – with the idea that the more which there is to say may vary from one discourse to another. If that is so, then the possibility lies

open that, while we are entitled to claim truth for both moral judgments, say, and statements about the physical properties of things, the kind of truth we (are entitled to) claim for the former is different from the kind we (are entitled to) claim for the latter, and that they differ in ways germane to R/AR disputes.

Cruces

It is a consequence of this last point that there is nothing in Wright's minimalist conceptions of truth and assertion which excludes – and it is clear that there is nothing which requires – that the notion of truth applicable within a given discourse is an evidentially unconstrained one, of the kind to which a Dummettian realist aspires. Here, then, is one crux at which realists and anti-realists may part company, whilst agreeing that the statements in dispute are minimally truth-apt, and that many of them are true. But a great part of the interest and importance of Wright's reconfiguration of the debate lies in its purported identification of a number of other R/AR cruces where neither protagonist is, or need be, committed to the possibility of evidence-transcendent truth. We can get them into a useful perspective by beginning with a few further remarks on the character of the disagreement between the Dummettian realist and her opponent.

This can be redescribed in terms of Wright's notion of *superassertibility*. Roughly, a statement is superassertible if it is warrantedly assertible and is, as a matter of fact, destined to remain so no matter how our state of information is improved. Less roughly, "A statement is superassertible … if and only if it is, or can be, warranted and some warrant for it would survive arbitrarily close scrutiny of its pedigree and arbitrarily extensive increments to or other forms of improvement of our information."[35] This notion, and the claims Wright makes for it, demand a much fuller discussion that can be given here; for present purposes, two points are crucial. The first is that – or so, anyway, Wright argues (1992, pp. 44–70) – the predicate "is superassertible" meets the conditions which the minimalist conception holds to be necessary and sufficient for a truth predicate. Reason to think a statement superassertible is reason to think it is warrantedly assertible and conversely, so that 'superassertible' and 'warrantedly assertible' coincide in normative force; but a statement may be warrantedly assertible and yet not superassertible (increments to our information may destroy our warrant), so that the two predicates may diverge in extension. The second is that superassertibility is essentially an evidentially constrained notion: a statement cannot be superassertible unless we can be warranted in taking it to be so. In virtue of the first point, superassertibility can, under certain assumptions, provide an interpretation or model of the truth predicate applicable within a given discourse. But, by the second point, if the truth predicate for a discourse admits of potentially evidence-transcendent applications, then it possesses a feature which superassertibility necessarily lacks and so cannot be so interpreted. A Dummettian R/AR dispute is, then, a dispute concerning the capacity of superassertibility to serve as an adequate interpretation of the truth predicate for the problematic region of discourse. The Dummettian realist contends that the truth predicate has a characteristic – that of potential evidence-transcendence – going beyond anything involved in satisfaction of the minimalist platitudes, which enforces a distinction between truth and superassertibility.

In this case, the additional feature of truth for which the realist contends is one which would require us to accept not only that truth and superassertibility are distinct *notions*, but that they diverge, at least potentially, in *extension*. But could there not, Wright now asks,

be supplementary characteristics of the truth predicate for a given discourse which demand a *conceptual* distinction between truth and superassertibility without, however, entailing (potential) divergence in their *extensions*? There should be, if Wright's approach is to assuage the misgivings which – as previously suggested – should lead us to doubt the capacity of Dummett's general characterization to do justice to the variety and richness of R/AR disputes. He contends that there are (at least) three such features, each representing ways in which the notion of objectivity for statements of a given discourse might be interpreted, and so apt to form the focus of R/AR disagreement without raising questions about evidence-transcendence.

One issue is whether the discourse satisfies the *Cognitive Command* constraint, which it will do just in case it is *a priori* that differences of opinion arising within it can be satisfactorily explained only in terms of something worth describing as a cognitive shortcoming in one or other of the disagreed parties.[36] It might, to illustrate with one of Wright's own examples, seem quite obvious that talk of what's funny, or beautiful, fails this test – comic and aesthetic tastes may simply differ; there will doubtless be a causal explanation why I find Buster Keaton hilarious, or Rubens's women beautiful, while you do not – but it need not be one that finds cognitive fault in either of us. Perhaps the same goes for moral judgment, but this is evidently more arguable, and so could be what's at stake between moral realists and their opponents. Rightly or wrongly, virtually all of us will think that stock market reports, summaries of football results, and rainfall records pass the test.[37]

Another – the *Euthyphro contrast* (cf. Wright, 1992, pp. 108–139; also Wright, 1988b) – concerns whether our (cognitively) best judgments in a given area are to be regarded as tracking an independently constituted realm of facts (the realist view); or whether, rather, we should view truth for the discourse's statements as somehow determined by, or constituted out of, our best judgments (the anti-realist option). The label is intended to recall Plato's dialogue, which has Plato maintaining that pious acts are thought to be so by the gods because those acts are pious, while Euthyphro contends for the opposed view, that pious acts are so because the gods take them to be so. Realist and anti-realist may be presumed to agree that there will be a coincidence between the facts of the matter and our judgments made under optimal conditions. The issue then concerns the direction of dependence: Are such judgments true because they match up with independently constituted facts, or are those facts themselves no more than a reflection of our best judgments?

How might it be resolved? Wright's plausible suggestion – developing an idea of Mark Johnston's[38] – is, roughly, that the latter view should prevail if it is knowable *a priori* that the coincidence should hold. Thus it might with some plausibility be held that the truth of judgments about the colors of things simply consists in their being the judgments suitably endowed perceivers would make under optimal conditions, on the ground that it is guaranteed *a priori* that best-color judgments co-vary with the color facts. By contrast, it might be argued that even if our best judgments about the shapes of two- or three-dimensional things match up with the corresponding shape facts, this is a contingent and *a posteriori* matter. As Wright emphasizes, the suggested test demands much refinement if it is to be acceptable. For example, it would award the verdict too easily to the anti-realist, say about color facts, if the conditions for optimal judgment were specified merely as 'whatever conditions are needed and sufficient to ensure that our color judgments are true.' Rather, it must be possible to provide a substantial and independent characterization, reflecting the detailed epistemology of such judgments. It must also be required that the *a priori* coincidence of best judgment with the facts is not independently guaranteed, simply by our conception of the relevant facts themselves; for example, we may conceive of pain and other sensations as

being such that a sincere and unconfused subject is immune to ignorance and error in her present-tense judgments, but would not wish to say, for this reason, that the facts are constituted out of best judgments.

The third issue focuses on *Wide Cosmological Role* (cf. Wright, 1992, pp. 174–201; Divers and Miller, 1995): Do the facts which true statements of a given kind record have a role to play in explanations of further facts of other kinds, beyond facts about our beliefs and other attitudes, and can they figure in such explanations other than as objects of those attitudes? It might be contended that while moral or modal beliefs, for example, are apt to figure in explanations of our actions, desires, and beliefs, moral or modal facts themselves exert no influence on other goings-on; in contrast, facts about the primary qualities of bodies, for example, exert causal influence in the world at large. To the extent that this is so, we might think that this justifies a kind of realism about facts of the latter kind which is unwarranted in regard to the former.

These quite programmatic suggestions prompt a whole host of questions. How are the various realism-relevant conditions – satisfaction of Cognitive Command, passing the Euthyphro test, possession of Wide Cosmological Role, capacity for evidence-transcendent truth – related to one another? Is the first the least that can be required for a species of realism, and a precondition for kinds of realism marked off by the others, as Wright suggests? More generally, do these various conditions admit of a linear ordering, corresponding to more or less robust forms of realism? Is the list complete, or are there other realism-relevant conditions to be discerned? How does the classification relate to R/AR disputes about the existence of entities of problematic kinds? Is the merely minimal truth of statements of arithmetic, for instance, sufficient for the existence of numbers (Wright, 1992, pp. 28–29) or should a more substantial kind of truth be demanded? These and other questions must be left for discussion elsewhere, and have indeed already attracted a good deal of critical attention. To conclude, I shall comment briefly on just three doubts Wright's work has provoked.

While Wright's proposal clearly provides a more inclusive framework for the location and pursuit of R/AR disagreements than Dummett's appears to do, it may be held that there remain forms of opposition to realism which elude it. In particular, it has been claimed that the species of irrealism enjoined by the meaning-skeptical argument advanced by Kripke's Wittgenstein does not fit Wright's agenda, on the ground that it involves rejecting even the application of the supposedly neutral notion of minimal truth to statements about the meanings of words (cf. Edwards, 1994, pp. 63–65). Minimal truth is a prescriptive or normative notion – this is what sets it apart from a merely disquotational one – and this, so it is claimed, requires that a distinction can be drawn between future applications of words which would be correctly judged to accord our present understanding, and those which would not. But that distinction, in this view, is precisely what Kripke's Wittgenstein argues to be vacuous; so he denies that there are even minimal facts about meaning. Even if this is the right way to understand Kripke's skeptic – so that his position falls below the lower bound, as it were, which Wright thinks realists and anti-realists alike should surpass – I doubt that it amounts to a very serious criticism of his approach. Wright is not committed to claiming that *every* self-styled anti-realist will in fact acknowledge the applicability of minimal truth to problematic discourses; he will agree that expressivists about morals, for example, have denied it. It is another question whether rejection of even minimal truth leaves a defensible position. In the present case, he argues that it does not – that semantic irrealism is inherently unstable, because it inflates into a self-defeating global irrealism.[39]

Wright says that any predicate satisfying the platitudes that define minimal truth ought to be recognized as a truth predicate, adding that we should thus be "at least in principle open to the possibility of a pluralist view of truth: there may be a variety of notions … which pass the test" (Wright, 1992, p. 25). This – together with several further remarks in the same vein – might be taken to evince sympathy with a view according to which 'true' is *ambiguous*, bearing one sense in application to a discourse satisfying Cognitive Command (CC), another when the facts recorded by the discourse enjoy Wide Cosmological Role (WCR), and so on. Such an ambiguity thesis, like the somewhat parallel doctrine that 'exists' is ambiguous, bearing different senses as applied to feelings, to tables and chairs, or to numbers and sets, may be held to have little to be said for it and much against, and it may, accordingly, be thought a serious objection to Wright's position if it involves its endorsement. Sainsbury (1996; see also Pettit, 1996, and Jackson, 1994) makes this point, but suggests that there is no need for Wright to be committed to it; just as what (mis)leads philosophers into postulating an ambiguity in 'exists' can be accommodated by acknowledging instead that feelings, tables, and numbers are different kinds of entity, so, he proposes, Wright is best understood as proposing, not that 'true' is ambiguous, but that different kinds of thing are involved in the truth (uniformly understood as minimal) of statements of different types. This seems to involve saying that moral states of affairs (supposing them to be recorded by statements satisfying at most CC) are of a different kind from, say, those recorded by statements about the primary qualities of bodies (assuming these to satisfy WCR). In support of this suggestion, Sainsbury observes that 'true' does not in fact figure in Wright's formulations of CC, WCR, and others. The attractions of this view are obvious, but it is not clear that it speaks adequately to the threatened objection. Minimalism involves the Correspondence Platitude (Wright, 1992, pp. 25–27) that a statement is true if and only if it corresponds to the facts: it is thus not clear that Wright could avoid regarding 'true' as ambiguous in Sainsbury's way, since a distinction between different kinds of fact will induce, via this platitude, a distinction between kinds of truth. There is also a certain cost: Wright sees himself as preserving what he takes to be sound in Dummett's characterization of R/AR disputes – that is, the idea that they concern the character of the truth predicate applicable within a problematic discourse – whilst freeing it from the exclusive focus on matters of evidence-transcendence and bivalence. But it does not seem possible to do this without retaining the idea that what the truth of statements consists in varies across different discourses. The question is whether that idea enforces an ambiguity thesis. The analogy sometimes drawn between the assertoric use of language and games may be helpful here. Perhaps, as Wright at one point claims, we may view the minimalist platitudes as encapsulating the essential core of a single notion of truth (Wright, 1992, p. 38), which may be filled out in different ways in relation to different discourses, somewhat as what constitutes winning may vary across different games, without inducing any ambiguity in the word 'win' (cf. Wright, 1996).

A third cause for concern, discussed by Wright himself and pressed by some critics (Sainsbury, 1996), is that CC may fail to amount to a significant constraint over and above the requirements – syntax and discipline – for merely minimal truth. When Wright first formulates CC, he expresses it as a constraint upon *explanation* of disagreements within the discourse:

> A discourse exhibits Cognitive Command if and only if it is a priori that differences of opinion arising within it can be satisfactorily explained only in terms of "divergent input" … or "unsuitable conditions" … or "malfunction." (Wright, 1992, pp. 92–93)

But he quickly falls into another formulation, in terms of the idea that any disagreement must "involve something worth describing as a *cognitive shortcoming*."[40] There is, of course, no harm done by employing the emphasized words as shorthand for the longer list of specific types of failing appearing in the original formulation. But the shift from the requirement that differences of opinion "can be satisfactorily explained only in terms of" cognitive shortcoming to the requirement that such differences must "involve" such shortcoming is not harmless. It invites a charge of trivialization which Wright himself confronts in this form:

> But it *is* a priori that any difference of opinion concerning the comic, when not attributable to vagueness and so on, must involve cognitive shortcoming, since, if all else fails, ignorance or error will at least be involved *concerning the truth value of the disputed statement*. (Wright, 1992, p. 149)

The charge troubles Wright, who expends much energy and ingenuity in an effort to meet it by arguing that the trivializing move requires defense of an "intuitional epistemology" invoking a "special faculty … apt for the production of non-inferentially justified beliefs essentially involving its [the discourse's] distinctive vocabulary," but that postulation of such a faculty ought to be constrained by considerations of best explanation which are not easily, and certainly not trivially, satisfied. But Wright has – or so it seems to me – conceded more than he need have done. He could have seen off the trivializer much more swiftly, by reverting to the opening formulation of CC. Suppose you and I are disagreed on some comic matter. Let it be granted that there is error – cognitive shortcoming, even – on one side or the other. Still, we are nowhere near to a satisfactory explanation, *in terms of cognitive shortcoming*,[41] of our disagreement, if all we have is that either I have come short, cognitively speaking, in thinking that p – my shortcoming consisting in the bare fact that I think that p when not-p – or you have come short in denying that p, yours consisting in the bare fact that you think that not-p when p. Appeal to this cannot *explain* our disagreement that I think that p when, as you think, it is the case that not-p is precisely *what* we want *explained*. Doubtless there is an explanation to be had why, say, I think that p when, as you think, not-p. If and when such an explanation is located, it may or may not comply with the requirements of CC – but there is so far no ground for thinking that it will, as the would-be trivializer requires.[42]

Notes

1 In a fuller discussion of the major influences on realist and anti-realist thought over the last 20 years or so, our discussion of Dummett's work would be balanced by an equally extensive examination of that of Hilary Putnam. The hard editorial decision to exclude direct discussion of Putnam's work from the present chapter was rendered somewhat easier by the inclusion of a separate chapter devoted to one of his central lines of argument (see Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM). Useful remarks about the relations between the positions of these two major figures in the debate are to be found in Putnam's introduction to the third volume of his philosophical papers (Putnam, 1983) and in Dummett (1994).

2 As Dummett himself emphasizes, it is important to distinguish between bivalence and the (putative) logical Law of Excluded Middle, which asserts validity of the schema 'A or not-A.' It is, in his considered view, the *semantic* principle of bivalence on which R/AR disputes turn, rather than the logical Law of Excluded Middle, which may be validated in other ways – by the adoption of a

supervaluational semantics, for example – which involve, as such, no distinctively realist commitment. In earlier writings, it is the latter which is taken to be at issue; but Dummett subsequently declares this to have been a mistake, shifting attention onto bivalence (cf. Dummett, 1978, p. xxx).

3 Dummett (1963, pp. 155–156) suggests that rejection of the Law of Excluded Middle need involve no departure from realism. But his views on the matter have shifted in respects going beyond those indicated in the preceding footnote (cf. 1982, p. 265; 1993b, pp. 467–468).

4 For formulations of this argument, see especially Dummett (1969, pp. 362–363) and Wright (1993a, p. 13). See also McGinn (1980, p. 26; 1981) and Tennant (1981; 1984).

5 For formulations of the argument, see Dummett (1976, pp. 79–83; 1973b, pp. 217, 224) and Wright (1993a, pp. 16ff, 53–54). In (1973b, pp. 216–218), Dummett gives three arguments – from communication, from knowledge of meaning, and from learning. The last two of these correspond fairly closely to the acquisition and manifestation arguments as described here. An excellent discussion of these arguments is given in Prawitz (1977).

6 For the original suggestion, see Dummett (1969, p. 363); see also Wright (1993a, pp. 89–90) and McGinn (1980, p. 26).

7 For useful remarks about the difficulty of extending the truth-value link gambit to quantification over an infinite domain, see Wright (1993a, pp. 89–90).

8 Cf. McDowell (1978, pp. 132–133); see also Wright (1993a, pp. 90–91).

9 Cf. McDowell (1978, pp. 135–136), McGinn (1980, p. 27), and Wright (1993a, essay 3, especially pp. 95 ff.).

10 A further difficulty concerns the appeal to the distinction between chronically and merely contingently evidence-transcendent truth. Wright (1993a, p. 14) gives a snappy argument to dispel the comfortable appearance that the former category is populated only by a handful of old chestnuts like those cited as examples. Let P be any statement that is contingently undetectably true. Then the statement that it is so will itself be not only true, but *chronically* undetectably true.

11 For discussion of this issue, see Dummett (1975) and Wright (1982).

12 Cf. Dummett (1976, pp. 98–101). Dummett objects that the suggestion 'fails to answer the question how we come to be able to assign to our sentences a meaning which is dependent upon a use to which we are unable to put them.' McGinn (1980, pp. 27–28), apparently taking himself to be rehearsing Dummett's objection, claims that the difficulty lies not in envisaging the required extension of our capacities, but in seeing how they might be manifested. This would make the acquisition challenge dependent on that based on manifestation. If I am right, the former challenge can be upheld without falling back on the manifestation argument, by appealing instead to the essentially normative character of meaning. Wright (1993a, pp. 23–26) develops an independent argument from normativity against semantic realism. Although McGinn agrees that none of the responses discussed thus far is effective against the acquisition challenge, he rejects it on the ground that it relies on an unacceptable reductionist assumption that 'no conception can enter into understanding a language that is not induced directly by sensorily presented conditions; any going beyond the observational must be impossible or arbitrary' (1980, pp. 28–29). I cannot discuss this objection fully here, but I do not think it can be right – McGinn gratuitously equates what can be recognized with what can be observationally verified, thereby rendering it utterly mysterious how it is that the anti-realist could possibly regard decidable arithmetic statements as unproblematic.

13 Cf. Currie and Eggenburger (1983, p. 271); Scruton (1976); see also Devitt (1983; 1984, ch. 12) for suggestions of the first line of response, and McGinn (1980, p. 30) for the second.

14 That is, beliefs having realist truth-conditions, as opposed to beliefs in the correctness of realism. The anti-realist will hardly deny that people may manifest realist beliefs in the latter sense.

15 Cf. Wright (1993a, p. 56) for a fuller statement of the difficulty. Among other attempts to meet the manifestation challenge by locating a distinctive explanatory advantage in the hypothesis of realist truth-conditions, perhaps the most impressive and rigorously developed is to be found in work by Christopher Peacocke. Regrettably, space does not permit discussion of it here. See

especially Peacocke (1986, chs 2 and 3; 1987; 1988; 1992). Wright (1993b) criticizes some of Peacocke's specific proposals.

16 It is not entirely uncontroversial that an anti-realist account of meaning must be revisionary of our inferential practice. Dummett has always argued that it would be (cf. also Tennant, 1987) but Wright (1981) defends the opposed view. See also his (1986a), which responds to Rasmussen and Ravnkilde (1982). Wright's considered view is that an anti-realist meaning theory will enforce revision (cf. 1992, ch. 2). In effect, there is a dilemma here for the realist who would appeal to actual inferential practice to meet the manifestation challenge: *either* semantic anti-realism must be revisionary of such practice, *or* it need not be; if so, then appeal to the practice begs the question by assuming that it is beyond criticism, as out of line with the kind of meaning we have assigned to our statements; if not, then the appeal falls flat, since the practice is not distinctively realist after all.

17 This criticism assumes that Strawson is seeking to make out an operational difference between realist and anti-realist practice. But it is possible that he has in mind a somewhat different, and potentially stronger, line of argument according to which, whilst nothing in our linguistic practice as such marks it out as realist, there are theoretical considerations which constrain us to view it as informed by realist conceptions. This line of thought is pursued in Edgington (1985) and Campbell (1994).

18 Among important recent work which it has not been possible to discuss here, I should mention Blackburn (1989) and Wright's reply (1989b). A sophisticated attempt to show that acceptance of the manifestation requirement, properly understood, is not inconsistent with the assignment of realist truth-conditions is to be found in Peacocke (1986).

19 Dummett (1993b, p. 465). This particularly useful paper summarizes, defends and in places qualifies the general approach first adopted in Dummett (1963) and developed in several papers over the intervening three decades.

20 The charge that Dummett mislocates 'the realism issue' is pressed by Devitt (1983; 1984, ch. 12). An excellent discussion of the charge is given in Taylor (1987).

21 For this line of reply, see Dummett (1993b, pp. 468–471). For extensive discussion of the contrast between reductionist and non-reductionist forms of anti-realism, and detailed argument for the claim that the latter affords the basis of a sustainable anti-realist challenge in areas besides mathematics, see Dummett (1963, pp. 156–165; 1982, especially pp. 239–263).

22 Mackie (1977, ch. 1); Field (1980, pp. 1–16; 1989, essays 1 and 2); Churchland (1979); Stich (1983); the special issue of *Mind & Language* 8(2), 1993, devoted to eliminativism contains several papers relevant to error-theoretical treatments of psychological discourse.

23 Dummett (1993b) draws a distinction between what he terms 'objectivist' and 'subjectivist' attitudes towards a class of 'apparent assertibles.' A subjectivist about moral utterances takes an expressivist view of them, contending that they serve to voice attitudes or feelings, rather than to make genuine statements. In contrast with this, Dummett says, he was all along concerned with R/AR disputes between parties both of which took an objectivist view of statements in the problematic class, adding that "the dispute between the subjectivist and the 'moral realist' is not one of those to which my comparative method was meant to apply: the issues in that dispute are different and prior to it" (Dummett, 1993b, p. 467). Doubtless Dummett is quite right that a significant disagreement about what notion of truth – one that is evidentially constrained or one that is not – has application within the problematic class will take place against the shared assumption that those statements are indeed truth-apt. The fact remains that subjectivism, in Dummett's sense, is one form which opposition to realism may assume, but one which his preferred characterization of R/AR disputes simply passes by. Indeed, it is hard not to read Dummett's remark as implicitly conceding as much.

24 Cf. Sorensen (1988, pp. 199–253; 1995); Williamson (1992; 1994a; 1994b); Cargile (1969; 1979, §36); R. Campbell (1974); and see Chapter 28, SORITES. The epistemicist conception is criticized in Hyde (1995) and also in Wright (1995), to which Sorensen (1995) replies.

25 Cf. Dummett (1993b, p. 468). Indeed, on what may be the most defensible version of the epistemicist view – see Williamson (1992) – it is not merely that we *do not* know the truth-values of vague statements – we *cannot* know them.

26 He writes: "An anti-realist may … [hold] that reality itself may be vague. Whereas, for the realist, vagueness inheres only in our forms of description" Dummett (1993b, p. 468). For a searching assessment of attempts to make the contrast in this way, see Sainsbury (1995).

27 More accurately, mathematical statements are never non-vacuously true – for whilst those which carry categorical existential commitments, like 'There exist prime numbers greater than 10^{17},' will be false, general laws like 'a + b = b + a' will, if construed as universally quantified material conditionals (e.g., "For all a, b: if a and b are numbers, then a + b = b + a"), be true, but merely vacuously, precisely because no objects satisfy their main antecedents.

28 A closely related anti-realist position in philosophy of science is van Fraassen's constructive empiricism, according to which good theoretical science aims, not at correct description of unobservable realities, but at empirical adequacy – that is, roughly, at maximizing derivability of correct observationally checkable conclusions from observationally verifiable premises. See van Fraassen (1980, ch. 1).

29 This necessarily simplified formulation of the objection corresponds closely to that given in Hale (1987, pp. 106–115). For more careful presentations, see Hale and Wright (1992; 1994). Field (1989, pp. 43–45) attempts to defuse the objection, and (1993) more fully, to which the last-cited paper by Hale and Wright replies.

30 Blackburn (1984, p. 171). Blackburn (1984, ch. 6) and several of the essays in his (1993) develop his program.

31 For a more detailed explanation, together with a sketch of what Blackburn takes to be the underlying logical form of expressive/evaluative compounds, see his (1984, pp. 189–196).

32 Cf. Wright (1988a, p. 33) and Hale (1992). Blackburn (1988) makes a significantly different attempt to preserve a projectivist construal of moral utterances. This later theory cannot be discussed here. It is criticized in detail in Hale (1992). See also Gibbard (1990) and Zangwill (1992).

33 Should this be construed as expressing disapproval of Henry's saying whatever it was without apologizing? Or as expressing approval for combining the belief that Henry said whatever it was with approval for his apologizing? The former seems to lose the distinction between the case where the speaker believes that Henry spoke offensively and that where she wants to leave that question open. The latter looks to run into trouble explaining why we don't assert conditionals with evaluative antecedents and factual consequents, like 'If Henry ought to apologize, he said that.' Both approaches appear committed to locating a hitherto unnoticed ambiguity in the conditional construction.

34 As Wright himself points out, Hilary Putnam's suggested equation of truth with warranted assertibility under ideal epistemic circumstances (cf. his 1981, p. 55) encounters a similar difficulty. But the issue is delicate – see Wright (1992, pp. 37–42).

35 Wright (1992, p. 48). The notion made its first appearance in his paper "Can a Davidsonian meaning-theory be construed in terms of assertibility?" Essay 14 in Wright (1993a, cf. esp. pp. 411–418).

36 Cf. Wright (1992, chs 2 and 3). A fuller specification of the constraint is given on pp. 92–93, where the notion of cognitive shortcoming is fleshed out – the disputants are not to be "working on the basis of different information (and hence guilty of ignorance or error …), or 'unsuitable conditions' (resulting in inattention or distraction and so in inferential error, or oversight of data and so on), or 'malfunction' (for example, prejudicial assessment of data, upwards or downwards, or dogma, or failings in other categories …)." Clearly this does not purport to be a finished list. Earlier, related formulations appear in Wright (1980; see pp. 448–449; 1986b; 1989a).

37 It merits emphasis – since statements from a discourse apt for merely minimal truth have to be 'disciplined' (i.e., subject to standards of correct assertion) – that Cognitive Command (CC) is a global constraint. *Any* disagreement over one of its statements must be traceable to cognitive

shortcoming, if a discourse is to satisfy CC; whereas while particular disagreements within a merely minimally truth-apt discourse may be rationally resolvable, cognitively blameless disagreement is not ruled out *a priori*. CC – or so Wright intends – is to take up the slack left by minimal truth-aptness.

38 The source was unpublished material. Johnston (1992, esp. appendix 3) gives an impression of his later views on the matter.

39 See Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §3, and further references given there.

40 Wright (1992, p. 93); cf. his reformulation of CC at p. 144.

41 There will, doubtless, be a causal explanation of some sort to be found, but that is beside the point.

42 Two caveats: (a) this suggestion assumes that Wright's later formulations of CC bring no advantage that is lost by reverting to the earlier ones; (b) there may be other ways to press the trivialization threat (cf. Sainsbury, 1996, and Williamson, 1994a) which cannot be defused by the simple move proposed here, and which call for a fuller discussion than space permits.

Thanks to Jim Edwards and Crispin Wright for very helpful comments.

References

Ayer, A. J. 1946. *Language, Truth and Logic*, 2nd edn. London: Victor Gollancz.

Blackburn, S. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Clarendon Press.

Blackburn, S. 1986. "Morals and modals." In *Fact, Science and Morality*, edited by G. Macdonald and C. Wright, pp. 119–142. Oxford: Blackwell.

Blackburn, S. 1988. "Attitudes and contents." *Ethics*, 98(3): 501–517.

Blackburn, S. 1989. "Manifesting realism." *Midwest Studies in Philosophy*, 14(1): 29–47.

Blackburn, S. 1993. *Essays in Quasi-Realism*. New York: Oxford University Press.

Campbell, J. 1994. *Past, Space, and Self*. Cambridge, MA: MIT Press.

Campbell, R. 1974. "The sorites paradox." *Philosophical Studies*, 26(3): 175–191.

Cargile, J. 1969. "The sorites paradox." *British Journal for the Philosophy of Science*, 20: 193–202.

Cargile, J. 1979. *Paradoxes*. Cambridge: Cambridge University Press.

Churchland, P. 1979. *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.

Currie, G., and P. Eggenburger. 1983. "Knowledge and meaning." *Noûs*, 17: 267–279.

Devitt, M. 1983. "Dummett's anti-realism." *Journal of Philosophy*, 80(2): 73–99.

Devitt, M. 1984. *Realism and Truth*. Oxford: Blackwell.

Divers, J., and A. Miller. 1995. "Minimalism and the unbearable lightness of being." *Philosophical Papers*, 24(2): 127–139.

Dummett, M. 1963. "Realism (1963)." First published in Dummett, 1978, pp. 145–165.

Dummett, M. 1969. "The reality of the past." *Proceedings of the Aristotelian Society*, 69(1): 239–258. Reprinted in Dummett, 1978, pp. 358–374.

Dummett, M. 1973a. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. 1973b. "The philosophical basis of intuitionistic logic." In *Logic Colloquium 1973*, edited by H. E. Rose and J. C. Shepherdson, pp. 5–40. Reprinted in Dummett, 1978, pp. 215–247.

Dummett, M. 1975. "Wang's paradox." *Synthese*, 30(3–4): 201–232. Reprinted in Dummett, 1978, pp. 248–268.

Dummett, M. 1976. "What is a theory of meaning? (2)" In *Truth and Meaning*, edited by G. Evans and J. McDowell, pp. 67–137. Oxford: Oxford University Press.

Dummett, M. 1978. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. 1982. "Realism." *Synthese*, 52(1): 55–112. Reprinted in Dummett, 1993b, pp. 230–276.

Dummett, M. 1991. *Frege: Philosophy of Mathematics*. London: Duckworth.

Dummett, M. 1993a. "Realism and antirealism." In Dummett, 1993b, pp. 464–478.

Dummett, M. 1993b. *The Seas of Language*. Oxford: Oxford University Press.
Dummett, M. 1994. "Wittgenstein on necessity." In *Reading Putnam*, edited by P. Clark and B. Hale, pp. 49–65. Oxford: Blackwell.
Edgington, D. 1985. "Verification and the manifestation of meaning." *Proceedings of the Aristotelian Society*, suppl. vol. 59: 33–52.
Edwards, J. 1994. "Debates about realism transposed into a new key: Critical notice of Crispin Wright, *Truth and Objectivity*." *Mind*, 103(409): 59–72.
Field, H. 1980. *Science without Numbers*. Oxford: Blackwell.
Field, H. 1989. *Realism, Mathematics and Modality*. Oxford: Blackwell.
Field, H. 1993. "The conceptual contingency of mathematical objects." *Mind*, 102(406): 285–299.
Frege, G. 1884. *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner. Translated into English by J. L. Austin as *The Foundations of Arithmetic*, rev. edn. Oxford: Blackwell, 1959.
Geach, P. 1965. "Assertion." *Philosophical Review*, 74(4): 449–465.
Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.
Haldane, J., and C. Wright, eds. 1992. *Reality, Representation and Projection*. Oxford: Oxford University Press.
Hale, B. 1987. *Abstract Objects*. Oxford: Blackwell.
Hale, B. 1992. "Can there be a logic of attitudes?" In Haldane and Wright, 1992, pp. 337–363.
Hale, B., and C. Wright. 1992. "Nominalism and the contingency of abstract objects." *Journal of Philosophy*, 89(3): 111–135.
Hale, B., and C. Wright. 1994. "A reductio ad surdum? Field on the contingency of mathematical objects." *Mind*, 103(410): 169–184.
Horwich, P. 1990. *Truth*. Oxford: Blackwell.
Hume, D. 1739. *A Treatise of Human Nature*, edited by L. A. Selby-Bigge. Oxford: Oxford University Press, 1888.
Hyde, D. 1995. "Review of T. Williamson, *Vagueness*." *Mind*, 104: 919–925.
Jackson, F. 1994. "Review of C. Wright, *Truth and Objectivity*." *Philosophical Books*, 35: 162–169.
Johnston, M. 1992. "Objectivity refigured." In Haldane and Wright, 1992, pp. 85–133.
Mackie, J. 1977. *Ethics – Inventing Right and Wrong*. London: Penguin.
McDowell, J. 1978. "On 'The reality of the past.'" In *Action and Interpretation*, edited by C. Hookway and P. Pettit, pp. 127–144. Cambridge: Cambridge University Press.
McGinn, C. 1980. "Truth and use." In Platts, 1980, pp. 19–40.
McGinn, C. 1981. "Reply to Tennant." *Analysis*, 41(3): 120–122.
Peacocke, C. 1986. *Thoughts: An Essay on Content*. Oxford: Blackwell.
Peacocke, C. 1987. "Understanding logical constants: a realist's account." *Proceedings of the British Academy*, 73: 153–200.
Peacocke, C. 1988. "The limits of intelligibility: a post-verificationist proposal." *Philosophical Review*, 97(4): 463–496.
Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.
Pettit, P. 1996. "Realism and truth: a comment on Crispin Wright *Truth and Objectivity*." Contribution to book symposium on Wright, 1992. In *Philosophy and Phenomenological Research*, 56(4): 883–890.
Platts, M., ed. 1980. *Reference, Truth and Reality*. London: Routledge and Kegan Paul.
Prawitz, D. 1977. "Meaning and proofs: on the conflict between classical and intuitionist logic." *Theoria*, 43(1): 2–40.
Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.
Putnam, H. 1983. *Realism and Reason: Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press.
Ramsey, F. P. 1927. "Facts and propositions." *Proceedings of the Aristotelian Society*, suppl. vol. 7: 153–170.
Rasmussen, S., and J. Ravnkilde. 1982. "Realism and logic." *Synthese*, 52(3): 379–437.
Sainsbury, R. M. 1995. "Why the world cannot be vague." *The Southern Journal of Philosophy*, suppl. vol. 33: 63–81.
Sainsbury, R. M. 1996. "Review: Crispin Wright, *Truth and Objectivity*." Contribution to book symposium on Wright, 1992. In *Philosophy and Phenomenological Research*, 56(4): 899–904.

Scruton, R. 1976. "Truth-conditions and criteria." *Proceedings of the Aristotelian Society*, suppl. vol. 50: 193–216.
Sorensen, R. 1988. *Blindspots*. Oxford: Oxford University Press.
Sorensen, R. 1995. "The epistemic conception of vagueness: comments on Wright." *The Southern Journal of Philosophy*, suppl. vol. 33: 161–170.
Stich, S. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.
Strawson, P. 1977. "Scruton and Wright on anti-realism, etc." *Proceedings of the Aristotelian Society*, 77: 15–22.
Taylor, B. 1987. "The truth in realism." *Revue Internationale de Philosophie*, 41(1): 45–63.
Tennant, N. 1981. "Is this a proof I see before me?" *Analysis*, 41(3): 115–19.
Tennant, N. 1984. "Were those disproofs I saw before me?" *Analysis*, 44(3): 97–105.
Tennant, N. 1987. *Anti-Realism and Logic*. Oxford: Oxford University Press.
van Fraassen, B. 1980. *The Scientific Image*. Oxford: Oxford University Press.
Williamson, T. 1992. "Vagueness and ignorance." In *Proceedings of the Aristotelian Society,* suppl. vol. 66: 145–162.
Williamson, T. 1994a. "A critical study of Crispin Wright, *Truth and Objectivity*." *International Journal of Philosophical Studies*, 2: 130–144.
Williamson, T.1994b. *Vagueness*. London: Routledge.
Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.
Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.
Wright, C. 1980: *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.
Wright, C. 1981. "Dummett and revisionism." *Philosophical Quarterly*, 31: 47–67.
Wright, C. 1982. "Strict finitism." *Synthese*, 51(2): 203–282. Reprinted in Wright, 1993a, pp. 107–175.
Wright, C. 1983. *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.
Wright, C. 1984. "Can a Davidsonian meaning-theory be construed in terms of assertibility?" In Wright, 1993a, pp. 403–428.
Wright, C. 1986a. "Realism, bivalence and classical logic." In Wright, 1993a, pp. 458–478.
Wright, C. 1986b. "Inventing logical necessity." In *Language, Mind and Logic*, edited by J. Butterfield, pp. 187–209. Cambridge: Cambridge University Press.
Wright, C. 1988a. "Realism, anti-realism, irrealism, quasi-realism." *Midwest Studies*, 12(1): 25–49.
Wright, C. 1988b. "Moral values, projection and secondary qualities." *Proceedings of the Aristotelian Society*, suppl. vol. 62: 1–26.
Wright, C. 1989a. "Necessity, caution and scepticism." *Proceedings of the Aristotelian Society*, suppl. vol. 63: 203–238.
Wright, C. 1989b. "Misconstruals made manifest." *Midwest Studies in Philosophy*, 14(1): 48–67.
Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.
Wright, C. 1993a. *Realism, Meaning and Truth*, 2nd edn. Oxford: Blackwell.
Wright, C. 1993b. "A note on two realist lines of argument." In Wright, 1993a, pp. 262–276.
Wright, C. 1995. "The epistemic conception of vagueness." *The Southern Journal of Philosophy*, suppl. vol. 33: 133–159.
Wright, C. 1996. Book précis and response to commentators (Horgan, Horwich, Pettit, Sainsbury, van Cleve, Williamson) for a book symposium on *Truth and Objectivity*. In *Philosophy and Phenomenological Research*, 54(4): 863–868 (précis) and 911–941 (responses).
Zangwill, N. 1992. "Moral modus ponens." *Ratio*, 5(2): 177–193.

Further Reading

Blackburn, S. 1980. "Truth, realism and the regulation of theory." *Midwest Studies in Philosophy*, 5(1): 353–372.

Blackburn, S. 1990. "Wittgenstein's irrealism." In *Wittgenstein: ein Neubewehrung*, edited by J. Brandl and R. Haller, pp. 13–26. Vienna: Hölder-Pinchler-Tempky.

Davidson, D. 1977. "Reality without reference." *Dialectica*, 31(3–4): 247–258. Reprinted in Davidson, 1984, pp. 215–225.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Divers, J., and A. Miller. 1994. "Rethinking realism: Critical notice of John Haldane and Crispin Wright (eds) *Reality, Representation and Projection*." *Mind*, 103(412): 519–533.

Dummett, M. 1959. "Truth." *Proceedings of the Aristotelian Society*, 59(1): 141–162. Reprinted, with additions, in Dummett, 1978, pp. 1–24.

Dummett, M. 1977. *Elements of Intuitionism*. Oxford: Oxford University Press.

Hale, B. 1986. "The compleat projectivist. (Critical notice of Blackburn 1984)." *Philosophical Quarterly*, 36: 65–84.

Horwich, P. 1982. "Three forms of realism." *Synthese*, 51(2): 181–202.

McDowell, J. 1981a. "Non-cognitivism and rule-following." In *Wittgenstein: To Follow a Rule*, edited by S. H. Holtzman and C. M. Leich, pp. 141–162. London: Routledge and Kegan Paul.

McDowell, J. 1981b. "Anti-realism and the epistemology of understanding." In *Meaning and Understanding*, edited by H. Parret and J. Bouveresse, pp. 225–248. Berlin, New York: De Gruyter.

McGinn, C. 1979. "An a priori argument for realism." *Journal of Philosophy*, 76(3): 113–133.

McGinn, C. 1982. "Realist semantics and content ascription." *Synthese*, 52(1): 113–134.

Peacocke, C. 1992. "Truth and proof." In Haldane and Wright, 1992, pp. 165–190.

Putnam, H. 1982. "Why there isn't a ready-made world." *Synthese*, 51(2): 141–168. Reprinted in Putnam, 1983, pp. 205–228.

Wiggins, D. 1976. "Truth, invention and the meaning of life." *Proceedings of the British Academy*, 62: 331–378.

Wright, C. 1994a. Reply to Jackson 1994. *Philosophical Books*, 35(3): 169–175.

Wright, C. 1994b. Reply to Williamson. *International Journal of Philosophical Studies*, 2: 327–341.

Postscript

BERNHARD WEISS

Bob Hale's chapter introduces one of Crispin Wright's realism-relevant cruces, the *Cognitive Command* constraint, as follows: a discourse will satisfy the constraint "just in case it is *a priori* that differences of opinion arising within it can be satisfactorily explained only in terms of something worth describing as a cognitive shortcoming in one or other of the disagreed parties" (this chapter, §5). Thus when a discourse fails the constraint it admits of disagreements in which the parties may be guilty of no cognitive shortcoming. Following recent literature let's call such disagreements 'faultless.'

Many writers argue that the notion of a faultless disagreement makes no sense; anything worth describing as a disagreement, they contend, must involve fault. Disputes, such as those in matters of taste or comic appreciation, are not disagreements, but simple divergences (see for instance, Boghossian, 2006). How the situation should be characterized largely depends on how we conceive of disagreement (see below). Others claim that the mere admission of minimal truth to a discourse suffices to establish that there is a mistake by one or other party; so all minimally truth-apt discourses would satisfy the constraint. Thus Wright's view would be incoherent because it requires us both to allow minimal truth to operate and to see it as a further and non-trivial question whether the discourse satisfies the Cognitive

Command constraint. Here's a version of the trivializing argument. Let's say that Tom and Dick differ over some matter apt for minimal truth. So Tom believes that P; and Dick believes that not-P. If Tom's belief is true then Dick believes something false. In which case Dick will have made a mistake, and will be at fault. So, assuming that no one has made a mistake, we must conclude that Tom's belief is not true. But then Tom has made a mistake, and is at fault, again contradicting the assumption that no one has made a mistake. But, since the contradiction has been derived from that assumption alone, we need to reject it: someone has made a mistake. Minimal truth precludes the possibility of appropriately faultless disagreement.

Some have questioned the framing of the constraint in terms of the notion of a *cognitive* shortcoming. If thinking about the nature of realism involves characterizing those areas in which our judgments arise (or ought to arise) from epistemic engagement with objective fact, then we are in search of a characterization of distinctively cognitive engagement. So we seem to have presupposed just what needs clarification. Wright (2003, essays 4 and 5) tries to finesse this point by arguing that any theorist claiming the presence of a cognitive shortcoming will then need to provide an epistemology to substantiate that claim. So the constraint works by imposing a distinctive obligation on the realist, distinguishing her from her anti-realist opponent. This line of thought helps somewhat with the trivializing maneuver because it is now possible to argue that anyone who accepts the trivializing argument must provide an epistemology for the relevant class of claims. And, where the judgments made are not inferred from others, this will involve a *sui generis* epistemology, for example, postulating an ability to intuit the fact of the comic matter. But, despite this, the argument still appears troubling because it seems independently plausible that to believe an untruth is to make a mistake. Thus, even if we can't give an epistemology which explains the source of the mistake, we are committed to there being one. Put differently, if the advocate of a discourse's sustaining Cognitive Command is committed, problematically, to an intuitional epistemology for it, then the argument seems to show that an advocate of minimal truth is likewise committed.

In order to make progress, the proponent of Cognitive Command needs to find a way to allow that neither Tom nor Dick is making a mistake, though they have contradictory views. One might relativize truth to something like a standard of taste. The trivializing argument then breaks down, since we can allow that Tom's belief is true relative to *his* standard of taste and Dick's belief is true relative to *his* standard of taste. So neither party need have made a mistake, where making a mistake is believing something false from the perspective of one's own standards. Relative truth has been fruitfully discussed in recent literature; itself provides a way of opposing realism; and might provide a way, not merely of explaining the possibility of faultless disagreement, but of cashing out notions of representation that Cognitive Command had aimed at. Thus we might say that a discourse fails to be fully representational, not if it fails the Cognitive Command constraint, but if it admits of a relativized notion of truth.

Relativizing Truth

Let's be clear: our focus is not relative truth as globally applicable but the idea that local regions of discourse may demand a relative notion of truth. We shall try to be a little more precise about relative truth and then we'll question whether it allows for a conception of faultless disagreement.

Familiarly, some features of language – such as indexicals and demonstratives – entail that the content of utterances of the same sentence may vary from context to context. To use Kaplan's (1989) terminology, the sentence has a fixed character, which can be employed in different contexts to express different contents. But applying the model here doesn't enable one to capture the phenomenon of faultless disagreement, since, though it allows for faultlessness, it expunges disagreement. Tom's utterance of 'Parsnips are delicious' would have a content roughly equivalent to his, Tom's, utterance of 'I find eating parsnips pleasurable'; and Dick's utterance of 'Parsnips are not delicious' equivalent to his, Dick's, utterance of 'I do not find eating parsnips pleasurable.' Of course, there is nothing contradictory about these utterances. Though Tom and Dick react differently to the experience of eating parsnips, this, in itself, doesn't constitute a disagreement. To cater for disagreement the tactic is, not to relativize content, but to relativize truth.[1] We can evaluate a content for truth at different circumstances, most obviously at different possible worlds. So the content expressed by my current utterance of 'I am hungry' is true in this world where I skipped breakfast, but is false in a world where I got up slightly earlier and ate a decent breakfast. The thought underlying relativism about truth is that we may be able to find other circumstances of evaluation – such as standards of taste – which allow the *same* content to have different truth-values at different such circumstances. So Tom's utterance is true at his circumstance of evaluation, since parsnips are delicious relative to his standard of taste, but is false at Dick's.

There are serious difficulties in making sense of a circumstance of evaluation and in explaining how a content is determined as true or false by a given circumstance of evaluation (see Boghossian, 2006). Wright (2006) suggests that superassertibility may provide a suitable model of truth here: we can imagine that 'Parsnips are delicious' is assertible (on the basis of Tom's subjective reaction) for Tom and is enduringly so, but 'Parsnips are not delicious' is assertible (on the basis of Dick's subjective reaction) for Dick and is enduringly so. But let us rather question whether the model provides a good way of understanding faultless disagreement.

It seems clear that we can make sense of the absence of fault – neither Tom nor Dick is at fault from *his* perspective – but do we have a clear disagreement?[2] We need to be clearer about what it is to disagree. MacFarlane plausibly writes:

> [T]wo parties disagree (as assessed from [circumstance] C) if … (a) there is a [content] that one party accepts and the other rejects, and (b) the acceptance and rejection *cannot* both be accurate (as assessed from C). (MacFarlane, 2007, p. 26)

Then we have an account of accuracy:

> An acceptance (rejection) of a content, p, is accurate as assessed from a circumstance C iff p is true (false) at the circumstance C.

When we put this together with the account of disagreement we get this simplification, since the same content cannot be both true and false at a single circumstance of evaluation:

> Two parties *disagree* just in case there is a content that one party accepts and the other rejects.

Take it that:

If an acceptance (rejection) is accurate then it is faultless.

And now it is possible that Tom can accurately accept the content *parsnips are delicious* and Dick can accurately reject that content. So they disagree but do so faultlessly.

Fine; but what sense do we have that the acceptance and rejection are in some kind of tension, a tension which betokens disagreement? There seem to be two possible sources of the tension. So, first, we might say that there is a tension between the two acts because the success of the one precludes the success of the other. But, second, we might say that there is a tension between the two acts because there is a conflict between fulfilling their respective commitments.

To be sure, there is no perspective, no circumstance of evaluation, in which what both Tom and Dick do are regarded as successful. So, no matter who Harry is, Harry will not be able rationally to accept that each act was successful. But why should either Tom or Dick care about success from Harry's perspective? Given that the *agent's* aims will determine success or failure of an act and that these are circumscribed by her perspective, no other perspective can have a bearing on the success of her act. The notion of aim is too agent-centered to do the right work, at least, in a relativistic setting.[3] Each speaker might well achieve her aim (and will do so if the disagreement is faultless).

So, from the perspective of the speaker, only her own circumstance of assessment is relevant to her aims. MacFarlane instead locates the relevance of *others*' assessments in the (Brandomian; see Brandom, 1994; 2001) business of giving and asking for reasons, of challenging assertions and responding to challenges. He takes it to be a brute fact about our practice that:

- one is entitled to challenge an assertion when one has good grounds for thinking that the assertion was not accurate (relative to the context of assessment one occupies in issuing the challenge), and
- a successful response to such a challenge consists in a demonstration that the assertion was, in fact, accurate (relative to the context of assessment one occupies in giving the response). (MacFarlane, 2007, p. 28)

So MacFarlane's strategy is to bring in circumstances of evaluation other than that of the speaker by focusing on the normative status of her assertion and on the commitments she incurs in asserting, namely, commitments to respond to legitimate challenges.

What drives the whole process is the insertion of acts of assertion in the business of giving and asking for reasons. But, when Dick challenges Tom, his challenge is bound to seem misplaced to Tom – because it emerges from an alien standard of taste – and when Tom responds his response is bound to seem misplaced – for analogous reasons. As MacFarlane admits, the parties may go on fruitlessly exchanging reasons until they get bored of the carry-on and move on.[4] Of course, in any exchange of reasons, parties may at times talk past one another, but when they do so we generally think of them as having failed in some manner. However, on the proffered account, this is just what we should expect: it seems an inevitable feature of the *practice* for which *practitioners* can in no way be blamed. Let us leave the debate at this inconclusive juncture; we still await a convincing conception of disagreement which allows for faultlessness. I want to close by considering a novel framework for realism/anti-realism debates.

Realism and Grounds

As Bob Hale outlines, Dummett understood disputes about realism as essentially semantic in nature and championed this conception as providing us with a neutral arena in which to prosecute the metaphysical debate. So the philosophy of language colonizes (an area of) metaphysics; but Kit Fine (2001) has recently put forward a view which resists these imperial ambitions, yet promises also to find a neutral way to negotiate the issue.

Fine thinks that realism is a view about the constitution of reality.[5] A primary notion implicated in the view is that of a ground. True propositions Q, R, and S ground another true proposition P just when it being that case that P consists in its being the case that Q, R, and S. As an example, Fine presents us with its being the case that Britain and Germany were at war in 1940 consists in…. (where the ellipsis is filled in with an exhaustive description of the relevant warlike activities of all relevant individuals).

So propositions Q, R, and S *actually* provide a ground for P, given the way the world is. Had things gone differently, P might have been grounded in a quite different set of propositions. So the grounding relation needs to be distinguished from a reduction of P to another set of propositions; in general, a reductive thesis holds that the meaning of the reduced propositions is given by the logical complex of reducing propositions. No such claim is involved in the grounding relation.

Fine's aim is to understand factuality in terms of grounding. He shows how a difference of view about the factuality of a proposition will play itself out in terms of facts about grounding, taken to be neutral in relation to the dispute. Although the details are complex, the basic idea is simple enough, though ingenious.[6] Dispute about the factuality of a proposition is pursued from that proposition to its grounds and ultimately to a dispute about the factuality of a constituent of a basic proposition. So, for instance, we might locate a dispute about the factuality of a moral proposition such as 'Harming dogs is wrong,' in the factuality of a constituent of this proposition, namely the property of being morally wrong ('wrongness' for short). So far there's no hint of a neutral forum for the debate, but Fine cleverly finds one by transferring the question about the factuality of wrongness into a question about the grounding of a proposition. The trick here is to note that wrongness will be a constituent of propositions which *both* sides agree are *factual*, for instance, the proposition that 'Harry believes that harming dogs is wrong' or that '"Wrong" refers to wrongness.' And now the question is: What grounds the truth of this factual proposition? The disputants must disagree, since they agree that a factual proposition cannot have grounds which essentially include a non-factual constituent.[7] So the factualist about morals can allow wrongness to be a constituent in the grounds of relevant factual propositions, the anti-factualist cannot. Thus their dispute about factuality must become a dispute about grounds, which, recall, Fine thinks is relatively uncontroversial.

We cannot examine the view properly here, but a concern is that the view begs metaphysical questions by helping itself to views about the nature and constitution of propositions.[8] Moreover, if, as some think, propositions need to be understood in terms of their linguistic expression, it is not clear that the proposal avoids thinking of realism in terms of the semantic relations between classes of statements.

The Variety of Conceptions of Realism

Dummett[9] sees realism from the start as a view about the relation of the world to our portrayal of it.[10] Fine demurs, seeing realism as a view about the constitution of reality. Though we might push these conceptions towards points of contact, would it impugn philosophical preoccupation with realism if we allow that there is a marked divergence here?[11] Not obviously so. The very same sense of a complex of issues – rather than a single crux – inhabits Dummett's later writings on the subject.[12] There he works still with the idea that the relevant notion of truth is crucial – because of its links with that of objectivity – and argues that theses about reference and about reduction are frequently important because of their impact on the notion of truth. However, he allows that these are pertinent beyond those consequences; evidently he does so, because he is concerned also to think about realism in terms of the constitution of the world.[13] Why expect that a complex of seemingly diverse debates should be capable of being distilled to a single underlying issue and otherwise rejected as simply confused?

Notes

1 The strategy is employed by Kölbel (2002; 2003), Lasersohn (2005), MacFarlane (2007; 2014), and Wright (2006).
2 It is worth noting that Wright gives up on talk about disagreement and talks instead of the parties as having contradictory beliefs. Rosenkranz (2008) also pushes a similar complaint.
3 As MacFarlane himself notes (2007, p. 27) as does Evans (1985).
4 Marques (2014) uses observations of this sort to argue that MacFarlane fails to ground a genuine notion of assertion.
5 A view shared with, among others, Devitt (see his 1991).
6 My presentation inevitably ignores many subtleties.
7 This is a simplification of Fine's principle (f): "any true imperfectly factual proposition [one including a non-factual constituent] has a perfectly factual ground [i.e., a ground including only factual constituents]" (2001, p. 18).
8 See Horwich (2010, essay 13) for a critique.
9 For recent debate of Dummett's views see Dummett (2004; 2006), Putnam (2010), and Dolev (2000).
10 Some don't see mind-independence in semantic terms; Khlentzos argues that even so, it should be seen as having semantic consequences (see his 2004, pp. 33–35).
11 As one recent commentator, Horwich (2006), claims.
12 See his paper "Realism" (1982) reprinted in his 1993.
13 Kitcher (2001) also sees these debates as having a dual character: (i) reduction of statements to a privileged class of statements; (ii) the applicability of correspondence truth to those statements. But he thinks anti-realists in each dispute are guilty of a single error.

References

Boghossian, P. 2006. "What is relativism?" In Greenough and Lynch, 2006, pp. 13–37.
Brandom, R. 1994. *Making it Explicit*. Cambridge, MA: Harvard University Press.
Brandom, R. 2001. *Articulating Reasons*. Cambridge, MA: Harvard University Press.
Devitt, M. 1991. *Realism and Truth*, 2nd edn. Princeton, NJ: Princeton University Press.
Dolev, Y. 2000. "Dummett's antirealism and time." *European Journal of Philosophy*, 8(3): 253–276.

Dummett, M. 1982. "Realism." *Synthese*, 52(1): 55–112. Reprinted in Dummett, 1993. pp. 230–276.
Dummett, M. 1993. *The Seas of Language*. Oxford: Oxford University Press.
Dummett, M. 2004. *Truth and the Past*. New York: Columbia University Press.
Dummett, M. 2006. *Thought and Reality*. Oxford: Clarendon Press.
Evans, G. 1985. "Does tense logic rest on a mistake?" In *his Collected Papers*, pp. 341–363. Oxford: Clarendon.
Fine, K. 2001. "The question of realism." *Philosopher's Imprint*, 1(1): 1–30.
Greenough, P. and M. Lynch, eds. 2006. *Realism and Truth*. Oxford: Oxford University Press.
Horwich, P. 2006. "A world without isms." In Greenough and Lynch, pp. 88–202.
Horwich, P. 2010 *Realism, Meaning and Truth*. Oxford: Oxford University Press.
Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Clarendon Press.
Khlentzos, D. 2004. *Naturalistic Realism and the Antirealist Challenge*. Cambridge, MA: MIT Press.
Kitcher, K. 2001. "Real realism: the Galilean strategy." *The Philosophical Review*, 110(2): 151–197.
Kölbel, M. 2002. *Truth Without Objectivity*. London: Routledge.
Kölbel, M. 2003. "Faultless disagreement." *Proceedings of the Aristotelian Society*, 104: 53–73.
Lasersohn, P. 2005. "Context dependence, disagreement and predicates of personal taste." *Linguistics and Philosophy*, 28(6): 643–686.
MacFarlane, J. 2007. "Relativism and disagreement." *Philosophical Studies*, 132(1): 17–31.
MacFarlane, J. 2014. *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press.
Marques, T. 2014. "Relative correctness." *Philosophical Studies*, 167(2): 361–373
Putnam, H. 2010. "Between Dolev and Dummett: some comments on 'antirealism, presentism and bivalence.'" *International Journal of Philosophical Studies*, 18(1): 91–96.
Rosenkranz, S. 2008. "Frege, relativism and faultless disagreement." In *Relative Truth*, edited by M. Garcia-Carpintero and M. Kölbel, pp. 225–239. Oxford: Oxford University Press.
Wright, C. 2003. *Saving the Differences*. Cambridge, MA: Harvard University Press.
Wright, C. 2006. "Intuitionism, realism, relativism and rhubarb." In Greenough and Lynch, 2006, pp. 38–60.

21

Theories of Truth

RALPH C. S. WALKER

1 Introduction: Problems with Correspondence

There are often said to be five main "theories of truth": the correspondence theory, the coherence theory, and the pragmatic, redundancy, and semantic theories. It is not really clear how far these theories are in competition with one another, for it is not clear how far they address the same question. However, they are all concerned with truth and falsity as properties of what people say or think. There are other uses of "truth" and "true," as when we speak of a true friend, but these are set aside, perhaps as derivative, at any rate as different.

Various views are held about how the content of what we say or think should be specified, and thus about what the bearers of truth are. Some people would specify it in terms of sentences, pieces of language, as uttered by a particular speaker at a particular time; for them, these would be the bearers of truth. Others would say that truth-bearers are statements or propositions, where these are thought of as what meaningful utterances of sentences express. Others again would say that they are judgments, mental contents which may or may not be expressible in language but which nevertheless embody thoughts. For many purposes this issue is a red herring; where it is not, I shall call attention to the fact. But in general, theories of truth have been concerned with a relation between the world and what we say or think about it; and much of the time it matters little whether the content of our thought is taken to be a judgment, a proposition, or just a sentence.

The correspondence theory of truth holds that for a judgment (or, say, a proposition) to be true is for it to correspond with the facts. In a sense this is obvious, but taken in that way the theory is unilluminating. Colloquially, "corresponds with the facts" can function as a long-winded way of saying "is true": so understood, the alleged theory becomes an empty tautology. To have content, it must at least claim that for a judgment to be true is for it to stand in a certain relationship ("correspondence") with something independent of that judgment, a fact or state of affairs in the world. Since this still seems hard to disagree with,

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

it is natural to think that a correspondence theory fully worthy of the name must go on to say something substantial about the relationship of correspondence, and also about the facts or states of affairs with which true judgments correspond. The paradigm of such a theory is Wittgenstein's in the *Tractatus*, and we shall return to it. It is not, however, clear that anything can be said about the correspondence relation, except that it is the relation in which a proposition stands to the world when it is true; nor is it clear that the relevant facts, or states of affairs, can be specified except as those which make a particular proposition true. These are standard objections to the correspondence theory. That they are plausible objections can be borne out by the odd position one can get into if one takes the correspondence theory, in this strong form, really seriously. Russell held such a view, in his Logical Atomist phase, and became much exercised over whether the world contains negative facts or not (Russell, 1956, p. 211). Is "The Queen is not bald" true in virtue of its correspondence with a special negative fact, or just in virtue of its lack of correspondence with any positive facts? For someone who held the correspondence theory in its strong form this would seem a real issue; others might think the theory had taken us off the rails.

2 The Coherence Theory and the Pragmatic Theory

The coherence theory of truth equates the truth of a judgment with its coherence with other beliefs. Different versions of the theory give different accounts of coherence, but in all its forms the point is to exhibit truth as an internal relation between beliefs. The theory holds that the truth, or falsity, of a belief can be determined by discovering whether or not it meets the appropriate test of coherence. In all its forms, again, the coherence in question is coherence with other things that are believed or subscribed to, or at least with other things that would be believed or subscribed to under specifiably ideal circumstances. The coherence theory is therefore not committed to the absurdity of accepting as true an arbitrary set of propositions which happens to be internally coherent. Whatever the standards of coherence may be, it seems likely that alternative sets of propositions will meet them: as Russell (1906–1907) pointed out, although the highly respectable Bishop Stubbs died in his bed, the proposition "Bishop Stubbs was hanged for murder" can readily be conjoined with a whole group of others to form a set which passes any plausible coherence test; and indeed, the same can be said of the propositions that make up any good work of realistic fiction. Russell thought this an objection to the coherence theory, but it is not, for the coherence theory is concerned with coherence not amongst arbitrary propositions, but amongst beliefs.

As to whose beliefs, different versions of the theory again hold different things. Some would equate them with the beliefs held by our society, or, perhaps, the beliefs held by humankind in general. Since the beliefs even of a single person will include some that are inconsistent, what is required cannot be coherence with all the beliefs in question but with a majority, or with a majority weighted in some way, perhaps in terms of how deeply they are held. Even so, we normally think that many of our most deeply held beliefs may turn out false, as beliefs in demons have done; and we also think that many of the truths to be discovered in years to come could not be shown to be true simply by their coherence with our present beliefs. Hence some coherence theorists regard truth not as coherence with what we actually do believe, but with what we would believe under idealized circumstances – perhaps, as Peirce (1878) suggested, at the end of all human enquiry. Anyone taking this line must have in mind some non-trivial specification of what these circumstances would be;

if they were just "the circumstances under which we would believe the truth" the theory would collapse into vacuity. Other coherence theorists, perhaps more prominent in the past than at present, have met the same difficulties by a different move, regarding truth as coherence not with human beliefs, actual or potential, but with the beliefs of God or of an Absolute Mind.

Coherence theorists differ, again, over how widely they extend their equation of truth with coherence. Some make it cover truth of all kinds; theirs could be called pure coherence theories of truth. Others extend it more narrowly, to cover only truth of a particular kind – moral truth, perhaps, or the truth of theoretical statements in science. Perhaps the commonest position is intermediate between these, and equates truth quite generally with coherence amongst beliefs, except for recognizing a special place for sense-experience. A plausible reading of Kant ascribes such a coherence theory to him, so far as the world of appearances is concerned. The coherence that is equated with truth is, then, a coherence not just amongst beliefs, but also with the deliverances of sense-experience, or, as Kant (1781–1787, A218/B266) puts it, "with the material conditions of experience, that is, with sensation." On such a view, certain beliefs will be directly supported by sense-experience, and the truth or falsity of these will be a matter of how well they fit with the experience. There will be others whose relation to sense-experience is less direct, and whose truth is a matter of coherence within the system of beliefs, including, of course, those beliefs directly related to experience. A pure coherence theory, in contrast, would say that even for those beliefs that seem most immediately experiential – "I seem to see something blue" – truth is a matter of coherence, coherence presumably with other beliefs that the subject has or others have at the same or later times (such as the belief that there was an ink bottle there, that people's impressions of their own experience are usually reliable, and so on).

The pragmatic theory of truth is akin to a coherence theory of this Kantian kind. It holds that the truth of a belief is a matter of whether it "works," that is, whether acting upon it pays off. Acting on it pays off just in case the experiences we have are those the belief led us to expect. Thus the pragmatic theory also makes truth a matter of coherence, but coherence with future experience (Peirce, 1878; James, 1907; Dewey, 1938; and cf. Misak, 1991). The Kantian theory takes this view of one class of beliefs, except that it does not give special weight to future experience over experience past or present. Pragmatists have often moved towards mixed theories of the Kantian type, both by giving equal weight to experience at any time, and by allowing that not all beliefs have their truth-values determined by experience directly: for some, truth is a matter of coherence amongst beliefs. Quine's position is a variant of this (Quine, 1969). For him there is a difference between the two types of belief, but one of degree only. Whether the resulting theory should be called a coherence theory or a pragmatic theory seems arbitrary.

3 Coherence and Correspondence

The correspondence theory, the coherence theory, and the pragmatic theory are often presented as alternatives only one of which can be true. We have seen something of how the coherence theory and the pragmatic theory relate; but there is no reason why a coherence theorist, or a pragmatist, should not accept the correspondence theory in any of the forms so far described. No sensible coherence theorist will deny that truths correspond with the facts, in that sense of "correspond with the facts" in which it is a synonym for "are true."

No coherence theorist need deny the uncontentious claim that for a judgment to be true is for it to stand in a certain relationship, which can be called correspondence, with some state of affairs in the world. As to the more disputable form of correspondence theory, which attempts to elucidate the relationship by saying something substantive and illuminating about the relation of correspondence and about facts or states of affairs, there is no reason why a coherence theorist should not subscribe to that as well – unless, of course, there are considerations against such a theory which render it untenable for anyone.

Why, then, are the coherence and correspondence theories taken to be incompatible? For that we need more precision about what the coherence theory claims. So far I have been describing coherence theorists as "equating" truth with coherence; but that is vague. The coherence theory of truth is the theory that truth is *constituted by* coherence amongst beliefs; likewise, the pragmatic theory of truth is the theory that truth is *constituted by* conformity to (usually future) experience. These theories seek to tell us what the truth of a judgment consists in, or, in other words, to exhibit the essential nature of truth. Neither theory is merely asserting a biconditional, nor even a necessary biconditional. It is uncontroversial that we use coherence amongst beliefs as a test of truth, and that we use conformity to experience in the same way. Without subscribing to either theory of truth, someone might well hold that a belief is true if and only if it passes one or other of these tests. We normally suppose that on the whole our beliefs reflect reality, and that those are true which cohere with experience and with the web of our other beliefs. Even if one believed that the biconditional held necessarily, one might still not be either a pragmatist or a coherence theorist of truth. One might, perhaps, hold that reality was itself coherent as a matter of metaphysical necessity, and that as a matter of the same necessity our beliefs about it are, on the whole, largely true. Something like this was the view of Bradley. Bradley is often said to have held a coherence theory of truth; but that is a mistake, and the mistake lies in the failure to see the difference between a theory that says truth consists in coherence, and one which only says that, of necessity, propositions are true if and only if they meet the coherence test. For Bradley the truth of a judgment consists in its matching reality, though because of the nature of reality a judgment can be true necessarily if and only if it satisfies the requirements of coherence (Bradley, 1893, chs 13–15 and 24). For the coherence theory of truth, on the other hand, it is in the satisfaction of those requirements that the truth of a judgment consists.

A claim about what truth consists in is not usually intended to be analytic. It is not supposed to be a claim about concepts, but rather to be about what that property, truth, really is. In this it resembles the claim that the heat of a body is the mean kinetic energy of its molecules. Neither claim is supposed to be established by considering what we ordinarily mean by our words, or by examining what is "contained in our concepts"; quite how they are to be established is a difficult question, of course, but establishing such things is part of the traditional task of metaphysics. Coherence theorists think philosophical reflection can show that there is nothing other than coherence for truth to amount to. Their claim is therefore that truth is the same property as coherence, but that to characterize it as coherence is more adequate, more illuminating of its nature.

The coherence theory of truth is thus a radical, and at first sight highly counter-intuitive, thesis. True propositions can be said to correspond to facts, but since the theory holds that truth consists in coherence, the facts themselves are not independent of this coherence that determines truth. Coherence determines what propositions are true, and therefore what the facts are. Hence the facts are determined by what is believed, or would be under specifiable

circumstances; not, of course, that whatever is believed is true, but the truth is what coheres with the main body of beliefs, and can be equated with what would be believed by someone who believed all and only those propositions that so cohere. This is not the view of common sense. The commonsense view is that facts are what they are independently of what anyone believes about them, and independently of what anyone would believe about them under idealized circumstances – unless the idealization were of the trivial kind (such as "under those circumstances in which people would believe what is true"). Truth is, then, a matter of beliefs or propositions matching this independent reality. That gives us another sense, importantly different from the ones noticed before, in which we can speak of a correspondence theory of truth: as a name for this commonsense view. In this sense, unlike the others, the correspondence theory is incompatible with the coherence theory. It asserts just what the coherence theory denies, namely that truth consists in a relation between the proposition in question and something in the world which makes it true, where this something is taken to obtain independently of what anyone believes (or would believe) about it. Like the coherence theory, then, it is a theory about what truth consists in.

The classical proponents of the pragmatic theory are Peirce and James, and it is sometimes difficult to be sure exactly what their theory is; but as I have implied, it (usually) appears also to be a theory about what truth consists in. If it is not, it collapses into the claim that it is by its consonance with (future) experience that we discover what is true; but it would be distinctly misleading to call that epistemological thesis a theory of truth – just as it would be misleading to give that title to the claim that truth is discovered by testing for coherence amongst our beliefs. To the extent to which they reject the coherence theory, pragmatists can accept that truth is a relation between a proposition and something that is independent of what anyone believes, but only because they think of it as constituted, not by beliefs, but by experiences, which they take to be independent of and prior to our beliefs about them. Actually, though, it would be difficult to maintain quite *generally* that the truth of a belief consists in its consonance with experience, because alternative theories – inconsistent with one another – can each predict exactly the same empirical consequences. (Thus the Ptolemaic hypothesis, or the flat-earth theory, can be made to yield all the same predictions as more usual views, with sufficient adjustments to assumptions elsewhere.) Hence the pragmatist must accept coherence as a determinant of truth as well as consonance with experience; coherence, for example, with our ideas about simplicity of theory becomes of central importance for Quine (1969; cf. Peirce, 1878). So the pragmatist's position will be very much the same as that of those coherence theorists who allow a special place for experience; and it will be similarly counter-intuitive. The commonsense view of the matter regards the facts as obtaining independently of anyone's experiences of them, just as it regards them as independent of anyone's beliefs about them. So when "the correspondence theory of truth" is used as a name for this commonsense view, we can take it as holding that truth consists in a relation between a proposition and something in the world that makes it true, where this something obtains independently of anyone's experiences of it and of anyone's beliefs about it.[1]

4 Why Pragmatic and Coherence Theories are Attractive

It is no objection to a philosophical position that it is counter-intuitive. Pragmatists and coherence theorists would say that much of the attraction of what I have called the commonsense view derives from failing to distinguish it from "the correspondence theory of

truth" in one of the other senses of that term. In any case, they adopt their position, not because it seems to be what we normally think, or part of the ordinary meaning of "true," but because they feel there are strong pressures requiring us to accept what they hold about the nature of truth. These pressures are from considerations about knowledge and from considerations about meaning. To a large extent they gain their force from the fact that the commonsense view seems to leave open an awkward possibility: that our thoughts and our beliefs should wholly fail to describe the world around us. This is the possibility raised by Descartes's idea of the *malin génie*. If the world – the facts or the states of affairs that determine the truth-values of our propositions – is wholly independent of our beliefs about it and our experiences of it, what assurance could we ever have that our beliefs are really true?

The obvious answer is that, although the world is independent of our experiences and beliefs, they are not independent of the world. They are caused by it, and this fact somehow enables us to get all the assurance we need. That is what empiricists standardly say; and the reply to them is that the assurance is not provided. For all they can say, the possibility of an alternative causal origin remains, in the *malin génie*. Nothing in the content of our belief or the character of our experience can ever rule this out, for any candidate could have been placed there by the *génie*. Yet to take this suggestion seriously seems absurd. Kant's response was to distinguish between the everyday world of appearances and the world of things in themselves. About the latter we can know nothing. About the former we can know all that we ordinarily think we know, and we know that no *malin génie* deceives us; for the way things are in the world of appearances is determined conjointly by the content of our experience and the *a priori* principles that govern all our awareness. Truth in the world of appearances is a matter of coherence with these principles and with given experience. The *malin génie* hypothesis can therefore be ruled out, so far as that world is concerned: it does not cohere. Many of Kant's successors, and particularly Hegel, followed a similar line of thought, but rejected the hypothesis of things in themselves as redundant, even vacuous. This left no room for the *malin génie* even in the realm of the unknowable, for there is no such realm. Truth consists in the internal coherence of a system of beliefs. More recently others, some with very different conceptions of the world from Hegel's, have felt drawn to the same conclusion.

Often this has been because people have thought the hypothesis of an unknowable reality to be unintelligible. The verificationists of the Vienna Circle considered that an assertoric sentence could be meaningful only if it were possible to verify it, or perhaps to falsify it. Since it is not possible conclusively to verify, or to falsify, a great many of the things that we commonly say, the present-day proponents of such views require something less: properly to understand an assertoric sentence is to know under what conditions it would be warranted to assert or deny it. They support this with the argument that it is only through the association with its assertibility conditions that the meaning of an assertoric sentence could be taught, and only through the same association that it is possible to discover whether somebody knows how to use it correctly. Language is an instrument of public communication, but if one sought to convey to someone an idea that transcended all possibility of verification, it would be impossible ever to have reason to think one had been rightly understood. This at least is the position of Dummett (1978) and Wright (1987), and perhaps Wittgenstein (see Chapter 4, MEANING, USE, VERIFICATION, and Chapter 20, REALISM AND ITS OPPOSITIONS, §2).

Neither Dummett nor Wright sees it as leading to the coherence theory of truth, but others have done. Neurath (1931; 1932–1933) and Putnam (1978; 1983) are perhaps the

clearest examples of those who have explicitly adopted coherence theories as the consequence of a verificationist line of thought. It is natural to think them right in seeing the connection. Even if some of our assertoric sentences just describe what is given in experience (something Neurath disputes), clearly most of them make claims which go beyond what is presented; and on a verificationist view of things, understanding such a sentence involves grasping a set of rules to the effect that this or that circumstance establishes its truth or renders it warrantedly assertible. Now if one thought of these rules as yielding results which might be correct or incorrect, through their relation to an independent reality, one could resist any kind of coherence theory. But the verificationist does not think of them in that way. It is the rules themselves which determine what is correct or incorrect, for it is the rules which determine what verification is. There is no verification-transcendent reality with which their results can be compared. It is true that for contemporary anti-realists a particular assertion, like "Betty is in pain at *t*," may first be warrantedly assertible (in the light of her pain behavior) and then cease to be so (when we discover she is being filmed for a TV commercial); but by their own showing there can be no truth of the matter independent of whatever results the procedures of verification yield. If there were, it would have to be something that transcended verification; the possibility would again be opened up that our methods of verification are wrong.

A similar line of argument is sometimes put by saying that our conceptual scheme is, after all, our own. Our concepts, and hence the rules for their application, are provided by ourselves. Not that they are the result of our voluntary choice, of course, but besides the conceptual scheme embodied in our language and our thought there might have been equally viable alternatives. Our scheme is satisfactory, and yields truth about the world, just because it is our scheme itself (or, to be exact, coherence within it) that determines what constitutes truth. On that showing, alternative conceptual schemes might determine an alternative kind of truth; but these we reject, for they are not ours. Putnam often argues in this way, and Quine also, though Quine stresses that it is not just the coherence internal to our conceptual scheme that constitutes truth, but coherence also with experience. Davidson's position develops out of Quine's, but it differs in two essential ways. He attacks Quine for assigning the role he does to experience – beliefs can cohere, or fail to cohere, with other beliefs, but not with experience; and though we have beliefs we call empirical, their truth is determined by coherence amongst beliefs and nothing else. He also rejects the idea of alternative conceptual schemes. Nothing other than our own could constitute a conceptual scheme, for a being who lacked what is essentially our own system of beliefs could not be accounted rational, and neither concepts, beliefs, nor the language to express them could properly be ascribed to it (Davidson, 1986; for discussion of Davidson's overall approach, see Chapter 2, MEANING AND TRUTH-CONDITIONS, and Chapter 13, RADICAL INTERPRETATION). Davidson's seems to be a neat, clear-cut example of the purest form of coherence theory, though he is no longer willing to call it this himself (Davidson, 1990a; 1990b).

5 Why the Coherence Theory Fails

Despite its initial strangeness, then, there are strong reasons to adopt some version of the coherence theory of truth: reasons which may, indeed, seem compelling. Nevertheless, I think the theory is untenable: it offers an account of what truth consists in, but it is an account which depends on taking for granted the conception of truth.

It claims that the truth of p consists in its coherence with a set of beliefs that are actually held (or would be held, in non-trivially specifiable circumstances). If that were not so – if it claimed that truth amounted simply to coherence within an arbitrary set of propositions – the theory would be open to Russell's objection about Bishop Stubbs. But what about the claim that a certain belief, b_1 is actually held? If we suppose it true, in what does *its* truth consist? It must consist in coherence, for such is the theory – coherence with the other beliefs that are also held. Evidently, though, the same applies to them as well. This means that the Bishop Stubbs objection recurs after all. We can easily denominate an arbitrary set of internally coherent propositions including "Bishop Stubbs was hanged for murder," such that for each proposition p_n in the set, "It is believed that p_n" coheres with the original set. But that does not make it true that p_n is really believed; nor does the coherence of "Bishop Stubbs was hanged for murder" with the set make it true that Bishop Stubbs was hanged for murder.

What the theory requires is that it should be a fact that certain things are believed, a fact that obtains in its own right and not in virtue of some further coherence. A pure coherence theory of truth, which holds that truth *always* consists in coherence, cannot accommodate this. Nor can the impure coherence theory of Kant and the pragmatists, for it makes an exception only for truths about the content of experience. These philosophers treat claims about what people believe as being determined true or false by coherence, for they regard ascriptions of belief as part of our publicly shared theory about people's psychological states, and as having the same sort of status that theoretical claims of any other kind do.

There is in any case something rather unsatisfactory about a coherence theory of truth that is not pure. It is bound to give us a dual conception of truth: in some cases truth consists in coherence, but not in others. It is natural enough to think that we find out about the truth in different ways in different areas – we find out about mathematical truth in one way, truth about the latest news in quite another. We might express that by saying there are different criteria for truth in different areas. But that does not make it natural to think of a dual, or a multiple, conception of what truth is. We can detect warmth by touch or by thermometer, but our conception of it remains univocal. It would seem a lot less misleading to say that what coherence determines is not truth, but something else, which we might call quasi-truth. But then they would have to say that it is only in a very limited sphere that we can claim truth at all. Most of the things that we say are not true or false at all, because truth is correspondence with an independent reality and there is nothing for what we say to correspond to. They are at best quasi-true, or perhaps "useful." Pragmatists have sometimes been prepared to say this; those who take an instrumentalist view of scientific theories have often been willing to say it about the theoretical statements of the sciences. Blackburn, who regards moral and modal "truth" in this way, calls himself a quasi-realist in those areas, and contends that the language (and the logic) of "truth" and "falsity" can be used in these spheres to indicate coherence or the lack of it. Arguably it can. The trouble is that it is misleading, if not positively perverse, to use the word "true" to mark two entirely different relationships. (Wright as well as Blackburn would dissent from what I have just said; they emphasize the similarities rather than the differences in the roles assigned to "true"; Blackburn, 1984, ch. 6; Wright, 1992, ch. 4.)

6 Frege on Defining Truth

It may be felt, however, that the objection which was offered against pure coherence theories is too quick. It sounds a bit like the objection which Frege made, against the possibility of defining truth in any way at all, and opinions have differed as to whether Frege's case is convincing.

> If one were to say "A representation is true if it agrees with reality," that would achieve nothing, for in order to apply it one would have to decide, in a given case, whether a representation agreed with reality, or in other words whether it were true that the representation agreed with reality. Thus what is defined must itself be presupposed. The same would apply to every explication of the form "A is true if it has these or those properties, or stands to this or that in such-and-such a relation." The question would always arise in the particular case whether it is true that A has these or those properties, or stands to this or that in such-and-such a relation. (Frege, 1969, pp. 139–140)

What Frege calls an attempt to define truth can, I think, fairly be equated with what I have called a theory of what truth consists in. His objection to the correspondence theory, when offered as an account of what truth consists in, is that though it claims the truth of *p* consists in *p*'s correspondence with the facts, it must also admit that whether or not *p* corresponds with the facts is a matter of whether it is *true* that *p* corresponds with the facts. Similarly, the coherence theory must admit that the question whether *p* coheres with the beliefs that are held is the same as the question whether "*p* coheres with the beliefs that are held" is itself true. As Frege says, this seems to presuppose the concept that was being defined; and anyone who objects to the coherence theory along these lines will say that the problem with it arises not only over determining what it is for a belief to be actually held (or belong to the appropriate set), but more immediately over determining when one thing coheres with another. For the question whether *p* coheres with *q* is the question whether it is true that *p* coheres with *q*, but the truth of "*p* coheres with *q*" must consist in its coherence with something, say *r*. Its coherence with *r* must itself consist in the coherence (say with *s*) of "'*p* coheres with *q*" coheres with *r*,' and so we are into a vicious regress, even if we set aside the problem of determining values for *q*, *r*, *s*, and so on.

But Frege is wrong. An important difference between the correspondence theory and the coherence theory is crucial here. The correspondence theory is a theory of what truth consists in, but not a theory of what facts consist in. It can take the obtaining of a fact as ultimate. It does not have to consist in anything else. The coherence theory, however, has to hold that whether or not a fact obtains is determined by whether or not a certain proposition (the proposition that says this fact obtains) coheres in the appropriate way. The coherence theory is a theory of facts as well as of truth. The correspondence theory is a theory only of truth.

Let us take the correspondence theory first. As an account of what truth consists in, it holds that the truth of *p* consists in a relationship of correspondence between *p* and the facts. It also holds that whether or not this relationship obtains is itself a fact. It does not consist in anything else. In particular, then, it does not consist in the correspondence of "*p* corresponds with the facts" to the facts. Certainly, if *p* does correspond with the facts, then the proposition "*p* corresponds to the facts" will itself correspond to the facts; indeed, its truth – the truth of that *proposition* – consists in that correspondence. But the *fact* that *p* corresponds to the facts is a fact in its own right. Hence although there is certainly a regress of a kind, there is nothing vicious about it, any more than there is anything vicious about the observation that if *p* is true, it is true that *p* is true, and true that it is true that *p* is true. The obtaining of the correspondence relation between *p* and the facts is all that is required for "*p* corresponds with the facts" to correspond with the facts, for what the fact that "*p* corresponds with the facts" has to match is just the fact that the original correspondence relation obtains.

The coherence theorist can try a similar reply. The truth of *p* consists in its coherence with the set of beliefs S; but whether or not *p* coheres with S is something that holds in its

own right and does not consist in anything further. Certainly if "*p* coheres with S" is true, its truth consists in its own coherence with S, and the truth of "'*p* coheres with S" coheres with S' consists in coherence again. But from this we no more get a regress that is vicious than we did with the correspondence theory.

This will not do. A coherence theorist might indeed hold that it is a matter of fact, in its own right, whether or not propositions cohere with one another; but to do so would already be to make it impossible for one's coherence theory to be of the pure kind – the kind that offers an account of what truth consists in quite generally. That is because a coherence theory is inevitably a theory of facts as well as of truth. What makes something a fact, for the coherence theorist, is that the corresponding proposition is true, that is, that it coheres in the appropriate way. If it were otherwise, facts and truth would come apart. The truth of "the cat is on the mat" would consist in its coherence; but the fact that the cat is on the mat would obtain, or not obtain, quite independently of this. To hold, then, that whether or not propositions cohere with one another is simply a matter of fact, and not of any further coherence, must be to hold that the truth of "*p* coheres with *q*" does not consist in coherence, and thus to hold to the coherence theory in at best an impure form.

Coherence theorists have not usually been prepared to qualify their position in this way. Their accounts of coherence have differed radically in detail, but the pressures which led them to the coherence theory in the first place have led them also to the view that whatever it is that constitutes coherence must itself be determined by the beliefs that are held. Under "beliefs" here one must include not only the beliefs that the appropriate subjects would adumbrate, but also the rules of inference on which they rely and which decide for them what arguments are good and what arguments are bad. These are as much part of our conceptual scheme as anything else; they are as subject to the manipulation of the *malin génie* as any other beliefs, if the *malin génie* has room to manipulate at all. Hence, in their view, the coherence of our system of beliefs, or of our conceptual scheme, is a property entirely internal to it, the standards of coherence being set by the system itself. That was the theory's apparent advantage: it made truth an entirely internal matter, not a matter of matching an independent reality which seemed beyond our reach, as we could not ensure that our beliefs (and our principles of inference) reflected it correctly. Such a theory we have seen to be untenable. That certain particular beliefs are held, or would be held, under non-trivially specifiable circumstances must simply be a fact, a fact which consists in nothing further and which cannot consist in coherence. If it did, the theory would require the truth of a proposition to consist in its coherence with the set of beliefs which are held; but the truth of the claim that those beliefs are held would have to consist simply in its coherence with that set; and that will not do, for too many alternative sets of beliefs would count as "held" by that criterion, including Russell's remark about the Bishop. To put it in Frege's fashion, the theory requires that "what is defined" (the notion of truth) "must itself be presupposed."

7 The Correspondence Theory

We have seen that Frege's objection does not touch the correspondence theory of truth. It may be felt, all the same, that the correspondence theory will hardly do, for three connected reasons. In the first place, unless it is spelt out a good deal further, it hardly seems to be a *theory* at all: it just says that a true judgment or proposition is one which matches the way things are in a world that is independent of our beliefs about it and our experience of it.

Second, to show that the pure coherence theory is untenable is not to deprive of their persuasiveness the arguments that led to it. The correspondence theory seems to make the world so independent of our thoughts about it that it renders *utterly mysterious* how we can succeed in knowing about it or even thinking about it, a mystery which is not dispelled by invoking the word "correspondence." Third, this correspondence would have to be something that we constantly aim at, for the aim of our assertions is truth, but we can hardly aim at it without knowing something more substantial about it. As Putnam (1978; 1983) points out, if we grant that there is a world that is independent of our thoughts about it in the way required, then, whatever it is like, there is bound to be a large number of relations which hold between it, or the elements of it, and the things that we think and say. If it contains infinitely many elements, Putnam shows that there must be alternative ways of mapping the world on to what we say or think, alternatives which are just as systematic as the relation of correspondence can be supposed to be. By what possible feat could we pick out one of these relationships and decide that *this* is the one that we intend when we talk about correspondence? For detailed discussion of this matter, see Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM.

There have been attempts to give the correspondence theory more content, to meet the first concern at least. Two have been particularly important: that of Wittgenstein's *Tractatus* and that of Austin. For Wittgenstein (1922) the correspondence is a structural isomorphism. Propositions are pictures of facts: to the elements of the proposition correspond the elements of the relevant fact, and the way the elements of a proposition – ultimately, names – are fitted together to form the proposition again corresponds to the way the elements of reality – objects – are fitted together to constitute the fact. This can seem promising: if there is an account to be given of correspondence, what else can it amount to but some such structural isomorphism? Unfortunately, however, it does not succeed. It explains one correspondence (of the proposition with the fact) in terms of others (names with objects, structure of proposition with structure of fact) but since these are unexplained – according to Wittgenstein, inexplicable in principle – no real gain has been made. We may even feel something has been lost, if we feel that Wittgenstein's account of the way propositions are structured is altogether too Procrustean (a conclusion he later reached himself). And the theory has no answer to Putnam's objection. If there is one systematic mapping between propositions and facts that meets Wittgenstein's constraints, there are bound to be other different mappings that meet the same constraints, so that it remains obscure how we manage to intend the right one, and mysterious how we can succeed in knowing about reality.

Austin's correspondence theory is rather different. Statements, he says, are related to the world by conventions of two kinds, demonstrative and descriptive. Descriptive conventions correlate words, as they are standardly used, with "the *types* of situation, thing, event, &c., to be found in the world." Demonstrative conventions correlate words, as they are used on the particular occasion of utterance, with "the *historic* situations, &c., to be found in the world"; they are the conventions that determine reference. Then:

> A statement is said to be true when the historic state of affairs to which it is correlated by the demonstrative conventions (the one to which it "refers") is of a type with which the sentence used in making it is correlated by the descriptive conventions. (Austin, 1950, p. 116)

Austin construes "historic" broadly; still, it is arguable that his account is limited, as an account of truth, by its restriction to statements about historic states of affairs. Setting

this aside, there is something unsatisfactory about his reliance on "convention," as Strawson (1965) pointed out. Conventions govern language; what they do is to determine what is being said. Certainly, one can distinguish those conventions which one learns when learning a sentence's linguistic meaning – roughly, its Fregean sense – from those which determine its reference on a particular occasion. Taken together, these determine what is being said on the particular occasion; they determine the proposition expressed. But we were concerned with the relation between the proposition and the fact (or "historic state of affairs"). All that Austin tells us about this is that it is true when the relevant state of affairs is as it is said to be. This is disappointing. And if we try to generalize it – to take account of the fact that not every statement picks out a "historic state of affairs" and says that it is of a certain type – we are left only with what Mackie (1973) calls the notion of simple truth: that the statement or proposition is true if, and only if, things are as it says they are.

It is indeed hard to see how the correspondence theory can contrive to say more than this. The idea that the structure of the proposition somehow reflects the structure of the fact is really the only suggestion that seems at all promising, and no doubt both Austin and Wittgenstein were trying to capture it in their accounts, despite the fact that they went about it very differently. But complex structures can be said to "reflect" one another in alternative ways, because there are different ways of mapping one on to the other; and another of Strawson's objections to Austin is pertinent against any such theory. Sentences clearly have structure; arguably propositions do as well, since a proposition involves several concepts of different types; but do facts? "Facts," Strawson says, "are what statements (when true) state" (Strawson, 1950, p. 136). He is not denying that there is something in the world that makes a statement true. But the articulation of the world into facts, with a structure that reflects the propositions or the sentences we use to describe it, is simply the result of our way of thinking about it. The structure of the proposition reflects the structure of the fact, because we ascribe to the fact the structure of the proposition. Of course the world does have a structure; as I look around I can see a variety of objects strewn around on the floor, and their spatial arrangement (for example) is perfectly objective. But when I say, "This shirt is white," the corresponding fact can be said to be structured only because we think of the proposition as structured, consisting of a referring subject term and a predicative expression. Otherwise there is nothing specially structured about my shirt's being white; it just is.

The necessary and sufficient condition for *p*'s being true is that things should be as *p* says they are. If we are left with so little, do we still have a "correspondence theory"? There is reason to say so, whether or not we take it in the way that renders it inconsistent with the coherence and pragmatic theories. If we do take it in that way, then a substantial part of the point of calling it a correspondence theory can just be that it does exclude those alternatives: it maintains that truth consists in matching a wholly independent reality, and not in coherence with beliefs or experiences. Three objections were earlier raised to it, as considered in that way, and the first – that it hardly deserves the name of a theory – can thus be dismissed: the feeling that more must be said about correspondence dissolves once it is apparent there is nothing more that can be said. The second and third can, I think, be dismissed as well, though perhaps more tentatively. The third objection was that there will be various different relationships, even systematic ones, between our words or thoughts and the world, and nothing to enable us to pick out one of them as the intended correspondence. But the intended relationship is just that which gives our words the meanings they

have, and our concepts their application: then the sentence or proposition is true if things are as it says. Thus the third objection turns into the second, that it is mysterious how we can succeed in thinking or knowing anything about an independent world. On these issues there is much to be said, and this is not the place to say it, except to observe that we do seem perfectly able to say things about an independent world and show them to be true; not, perhaps, by establishing them so securely as altogether to rule out any possibility of deception by a *malin génie*, but then perhaps such a "thin and so to speak metaphysical" doubt (Descartes, 1964–1976, §vii, p. 36) need not worry us excessively.

If we take it in the form in which it is *not* inconsistent with the coherence and pragmatic theories, then certainly the "correspondence theory" tells us nothing startling, now that we have seen that the correspondence relation cannot be informatively elucidated. In this mild form it seems uncontroversial: it just tells us that truth is a relationship between what is said or thought and some fact or state of affairs in the world, namely, the relationship that obtains when things are as they are said or thought to be. It is not committed to any ontology of facts, or any account of their structure. It is thus free from the difficulty we noticed right at the outset, of being forced to postulate vast numbers of negative and hypothetical facts for negative and hypothetical propositions to correspond to, or else to give an alternative account of their truth. They also are true when things are as they say. The theory is still however worth stating – worth, even, calling a "theory" – because it appears inconsistent with the redundancy theory, and because on further investigation it turns out not to be quite as uncontroversial as it looks.

8 The Redundancy Theory

The conflict with the redundancy theory may be more apparent than real. For the redundancy theory (Ramsey, 1927) is not a theory of what truth consists in, but a theory about the meaning of the words "is true." It holds that "… is true" can be deleted without loss. One could combine this with the thesis that truth *consists in* a relation of correspondence, because to say that is not necessarily to say anything about what we mean when we use the expression "… is true." Correspondence theorists have usually been concerned with what truth consists in, not with the analysis of meanings. There have, however, been exceptions. Austin was one. His account, cleared of the confusion over conventions, reduced to the thesis that a statement or proposition is true if, and only if, things are as it says they are. But Austin would take this, very plausibly, as giving an analysis of what we mean when we say "'Socrates is wise" is true': things are as "Socrates is wise" says they are. This is different from an analysis which allows us just to delete "… is true" and forget about it.

The redundancy theorist's alternative analysis also has some initial plausibility so long as we consider only examples of the form "*p* is true," where a value for *p* is specified; or of the form "*p* is false," which the redundancy theorist will analyze as "Not-p." Even there, it arouses two sorts of reservation. One is that "*p* is true" does seem to differ, at least in force, from "*p*," adding confirmation or endorsement or something of the sort. The second is that – as Austin claimed – it seems to say something *about p* instead of just asserting it. Wright finds in the idea lying behind these reservations material for a deeper objection: the norms that govern "*p* is true" are, he argues, inevitably different from those that govern the assertion of *p* (Wright, 1992, ch. 1). In any case, though, the theory seems to require amendment (or at least development) as soon as we move on to examples of other kinds,

as Ramsey was himself aware. Theories which try to preserve the spirit of the redundancy theory while making such amendments are often called deflationary theories.

What about the self-referential "This statement is true" and "This statement is false"? Here "is true" can hardly just be dropped out, nor can "is false" simply be replaced by a negation. Redundancy theorists have often viewed this as an *advantage* of their analysis, since it makes out such paradox-involving sentences to be incoherent. The point is moot. It is one thing to design an artificial language in which paradox-involving sentences are not well formed, and another to claim that this is true of an ordinary language like ours. The paradoxical character of "This statement is false" is not to be denied, for it can readily be understood by anyone, and a theory which argues it away cannot claim accuracy as an analysis of what we mean. This matter, however, deserves a fuller discussion than it can receive here.

At any rate, the redundancy theory cannot dismiss as ill-formed "The first statement on page 36 is true" or "Whatever the Pope says is true"; nor, indeed, have redundancy theorists ever wished to. They analyze the first as something like "There is some value of p such that p is the first statement on page 36, and p," the second as "For all values of p, if the Pope says that p, then p." If we put this more colloquially we get "Things are as the first statement on page 36 says they are" and "Things are as the Pope says they are." Here, then, the redundancy theory gives the same analysis as the amended Austinian theory.

People have sometimes expressed unease about such formulations. Often they suggest that there is something fishy about the quantification. In such examples the quantified variable occurs twice, but it is not clear that it functions in the same way both times. In its first occurrence – "If the Pope says that p" – it is arguable that it stands in for a name: or more exactly, that "that p" stands in for a name, the name of a proposition. "That Socrates is wise," or "the proposition that Socrates is wise," designates the proposition that Socrates is wise, and so in general. Indeed, if we express the point in terms of sentences instead of propositions, the matter seems clear-cut: the if-clause becomes 'if the Pope says "p,"' and the readiest device for forming a name of a sentence is to put it between quotation marks. Philosophers are very familiar with quantifiers whose bound variables stand in for names, so this first occurrence of the variable looks unproblematic. But the second cannot be handled in the same way. If it were, we should have to replace p again by a name – a sentence in quotes, or a noun phrase referring to a proposition. In that case the then-clause would not be grammatically complete: to make it so we should have to add "is true." This would remove any residual space for the thought that "is true" is in some sense redundant. It would also amount to abandoning any attempt to characterize the truth relation. The thesis, that truth is the relation between a proposition (or a sentence) and the world when things are as it says they are, would turn into the thesis that truth is the relation which stands between a proposition (or sentence) and the world when the proposition (or sentence) is true.

We could try to resolve the problem by using a different kind of quantification. We might construe the quantifiers substitutionally, or we might reflect that just as second-order quantification seems possible in which the variables are replaced by predicative expressions, so there is room for a kind of quantification in which variables are replaced by propositional or sentential expressions (cf. Grover, Camp, and Belnap, 1975/Grover, 1992, ch. 3; see also Grover, 1992, ch. 1). What is really at issue, however, is whether the variable is functioning the same way in both cases. This raises complex problems, but here it will perhaps do to make two observations. One is that if we express the matter in terms of sentences rather than propositions, the variable does appear to function in two ways, so that we need an

account of how the two occurrences relate. The other is that if we express it in terms of propositions, it is at least arguable that the variable functions in one way only.

In fact, this seems a highly plausible claim. In the antecedent it is not "*p*" but "that *p*" which looks as if it might name a proposition. The "*p*" itself does not name a proposition but expresses it, just as the "*p*" in the consequent does. In neither case is the proposition asserted, but it is characteristic of propositions that they can occur in just such contexts as these. Propositions are not objects, any more than properties are, and to assimilate the occurrence of propositional expressions to that of names would be no more sensible than to assimilate predicates to names (cf. Prior, 1971, ch. 3; for a different account – perhaps less different than it looks – see Horwich, 1990, especially pp. 18–21). Confusion is encouraged here by persistence in asking the question, "What then *are* propositions?" with the implied demand that they should turn out to be items like sentences, or sets of sentences, or the ghostly inhabitants of Frege's Third Realm. Propositions are not sentences, but what sentences express; conditions can be specified under which two sentences express the same proposition; but beyond that there is nothing more to be said. (In the same way one can provide conditions for property-identity, but one cannot go further than that in giving a non-circular answer to the question "But what *are* properties?") There is no gap between understanding the proposition that Socrates is wise and knowing what it would be for Socrates to be wise, so it would be impossible to know that the Pope had expressed that proposition without also knowing what it would be for things to be as it says they are. With sentences it is (perhaps) a different matter: one might know that the Pope had uttered a certain sentence ("Socrates is wise") without knowing anything about what it would be for things to be that way, because one might know he had uttered that sentence but not know what it meant.

Philosophers have often been reluctant to talk about propositions, because of puzzlement over what they are, and over their identity-conditions (see Chapter 39, OBJECTS AND CRITERIA OF IDENTITY; also Chapter 14, PROPOSITIONAL ATTITUDES). They have felt that clarity can only be attained by analyzing claims that might appear to be about propositions in such a way that they turn out to be about sentences. If one takes this view, one will feel constrained to adopt the first alternative above, and construe the quantification as really being over sentences. It then seems clear that the two occurrences of the variable do function differently. It is natural, if not unavoidable, to construe "If the Pope says 'Socrates is wise' then Socrates is wise," as containing a reference to a sentence in the antecedent, and then as using that same sentence (and not referring to it) in the consequent. A "sentence" here is of course more than just a string of words; it is a grammatical sentence as uttered on a particular occasion. But to explain the connection between the quoted sentence and the sentence itself, as it occurs in the consequent, we shall have to add an extra clause. The most natural way to do this would be by adding that the sentence "Socrates is wise" means that Socrates is wise: "If the Pope says 'Socrates is wise,' and if 'Socrates is wise' means that Socrates is wise, then Socrates is wise." But anyone who rejected the alternative analysis, in which the quantification was taken to be over propositions, would refuse to accept this as adequate on similar grounds. We need an account of the relation between the noun phrase "that Socrates is wise," which in the newly added clause refers to a proposition, and the "Socrates is wise" of the consequent. Such a person would make a parallel objection to any alternative to "means" which still made use of a that-clause. The only possibility, therefore, if the connection is to be expressible at all, is to use the relationship "is true if and only if." If the Pope says "Socrates is wise," and if "Socrates is wise" is true iff Socrates is wise, then Socrates is wise.

This is unfortunate. The redundancy theory held that the meaning of "Whatever the Pope says is true" is captured by "Things are as the Pope says they are"; the apparently undemanding and inoffensive correspondence theory held the same, at least when construed as an account of the meaning of "is true." But whether proposed as an account of the meaning, or as a theory of what truth consists in, it will collapse into circularity if it must, itself, make use of the notion of truth. We cannot explicate truth or "is true" by rendering "Whatever the Pope says is true" as "For every value of *s*, and every value of *p*, if the Pope says *s*, and *s* is true iff *p*, then *p*."

9 The Semantic Theory

Tarski's (1935; 1944) "semantic conception of truth" offers a possible way out of this. What Tarski shows is that for a certain rather restricted class of languages, a non-circular explication can be given of "*s* is true iff *p*," for every sentence *s* of the language concerned. It is done by providing a finite set of axioms for the language, and a set of derivation rules which permit one to deduce from the axioms a theorem of the required form – "*s* is true iff *p*" – for each sentence *s* of the language. Here *s* designates a sentence of the language under examination; it does this in terms of its structure, exhibiting it as a concatenation of simpler object-language expressions. The proof itself (and the axioms and theorems) belong to a different language, a metalanguage in which the other language is described. Tarski attached great importance to this distinction between metalanguage and object language, because he was anxious to avoid the problems raised by paradoxes like "This sentence is false." The languages he is concerned with do not permit semantic predicates like "is true" and "is false" to be meaningfully applied, within a given language, to the sentences of that language itself: only (if at all) to the sentences of some other language, for which that language is a metalanguage.

This restriction is enough to prevent Tarski's ideas from applying to ordinary languages in an unqualified form, for clearly these do allow the sort of predication he ruled out. However, it might be possible to carry over his central idea, in some way, into an explication of "is true iff" which would work for an ordinary language. The central idea is that the relationship "is true iff" is fully characterized, for the language concerned, by the axiomatic theory which permits the derivation of theorems of the form "*s* is true iff *p*" for every sentence *s* of the language. The theory does not require a prior understanding of the notion of truth. For that reason, as Tarski sets it out, it does not use the word "true" at all, but rather generates sentences of the form "*s* is T iff *p*," which are often called T-sentences. This relationship – "is T iff" – is then claimed to match our ordinary conception of truth.

The derivation rules that the theory needs are the ordinary rules of deductive logic. The axioms match each of the primitive terms of the object language with a thing or set of things in the world, thus matching the term "Socrates" with Socrates and the term "is wise" with all those things which are wise; they also include clauses which make use of the structure of the object language to enable the derivation of T-sentences. Thus where "*n*" is a name in the object language and "is F" is a one-place predicate, "*n* is F" is T iff the item that is matched with "*n*" is amongst those which are matched with "is F": from which we can infer that "Socrates is wise" is T iff Socrates is wise. Again, "*n* is F and G" is T iff the item which matches *n* is amongst those matched with "is F" and also amongst those matched with "is G," so that "Socrates is wise and sober" is T iff Socrates is wise and Socrates is sober.

The relation of matching which the axioms implicitly define is called by Tarski "satisfaction." The satisfaction conditions become much more complex, of course, when one deals with many-place predicates and sentences of more elaborate structure, and the ingenuity of Tarski's account lies in how he deals with the complexity; but the main idea is the same.

Although Tarski says his account "will explain the meaning" of "truth" (Tarski, 1944, p. 351), he is not, of course, offering a conceptual analysis, but an account of what truth consists in. I suggested that it might be helpful to a correspondence theorist who thought it would not do, as it stood, to say the correspondence relation is the relation which holds between a proposition and the world when things are as the proposition says they are, because this involves an improper quantification over propositions. It is now clear that if the correspondence theorist seeks to give an *analysis* of "is true," Tarski's work will not in fact help, because it does not contribute to an analysis. It will not help the redundancy theorist either, since the redundancy theorist also seeks to give an analysis. But to someone who holds a correspondence theory of what truth consists in, it may seem to provide just what is needed. For that reason Tarski's account has sometimes itself been called a correspondence theory.

But does it really provide what is needed? It is more like a counsel of despair. We might hope an account of what truth consists in would tell us something about the nature of truth in general, but on Tarski's view truth does not consist in anything general. His account allows us only to say that truth is a property which characterizes "The cat is on the mat" iff the cat is on the mat, "Socrates is wise" iff Socrates is wise, and so on for all the sentences of the language under discussion. Indeed, for Tarski truth could perfectly well be characterized by a long list of this kind, if the object language were so limited as only to contain a limited number of sentences. The detour through structure and satisfaction is needed only because we want to deal with languages which include indefinitely many sentences, and therefore need an account which has a limited number of axioms but yields an indefinite number of T-sentences. But it tells us nothing *general* about truth at all. A Tarskian truth-theory is a theory for a particular language, enabling us to derive T-sentences for all the sentences of that language. It is the satisfaction axioms for the language L_0 which determine what constitutes truth in L_0; the satisfaction axioms for L_1 which determine what constitutes truth in L_1; and so on. This gives us no general account of what truth consists in, and so it does not give us what we were looking for. A theory of what truth consists in was meant to be an account of what that property, truth, really is: not just truth-in-L_0, or truth-in-L_6. In fact, not only does he not give us what we wanted, Tarski goes further and assures us that we *cannot have* what we wanted. No general account of truth is possible, because it would have to be an account of truth-in-all-languages, including the language in which it is itself expressed. But a language capable of giving an account of "is true" for itself would have to be a language in which Tarski's rigorous distinction between object language and metalanguage was obliterated: a language in which sentences like "This sentence is true" are well-formed, and hence also "This sentence is false."

As already remarked, natural languages just do seem to violate the object-language/metalanguage distinction, and to give meaning to such sentences. Some people have therefore sought to retain Tarski's insights about the relation between truth and satisfaction, while abandoning that distinction and proposing alternative ways of handling the paradoxes. Kripke (1975) offers a particularly interesting attempt at this. However, any such account can still only offer us theories of truth for particular languages, and remains powerless to say anything general about what truth consists in. This is because, like Tarski, they

still explicate truth by providing a method of specifying for particular sentences of the language the circumstances in which they are to be called true: truth is the property which belongs to "The cat is on the mat" iff the cat is on the mat, to "Snow is white" iff snow is white, and so on. We could achieve the generality we are looking for only by a more radical move. Instead of introducing satisfaction as implicitly defined by a set of axioms relating the expressions of a particular language with sets of things in the world, we might just say that the satisfaction conditions relate a predicate to the things it applies to, a name to the thing it designates, and so on, and that they therefore yield an account of truth for any language by allowing the derivation of T-sentences for the sentence of that language. But *that* would bring us back to something very familiar. For simple unqualified sentences, what it tells us is that truth is the property a sentence has iff the items it designates have the properties it says they do. More generally, it tells us that truth is the property a sentence has iff things are as the sentence says they are. To say that, of course, is to say something which certain philosophers find objectionable, because of the implicit quantification over propositions: for every sentence *s* and for every proposition *p*, if *s* says that *p*, then *p*.

So we are faced with an alternative. Either we must accept that the correspondence relation can be legitimately, if not very excitingly, expressed by some such formulation as "Things are as *s* says they are," or else we must conclude, with Tarski, that nothing general can be said about truth at all: to the question "What is truth?" there is nothing to be said. At least this would have the merit of explaining why Pilate did not get an answer.

Note

1 Arguably there is room for an intermediate position here – one occupied, according to Wright (1992, ch. 3; 1995), by the semantic anti-realism explored by Dummett, Putnam, and himself. It too holds that truth consists in a relation between a proposition and something in the world that makes it true. It denies the coherence theorist's claim that the world is constituted only by coherence amongst beliefs, and holds instead that the world exists independently of what anyone believes, or would believe, about it. However, it maintains also that a statement can represent or describe the world only to the extent that there can, at least in principle, be grounds that would warrant its assertion. And it is only in so far as it can represent or describe the world that a statement is a candidate for truth. Hence, if we use the term "facts" for those aspects of reality in virtue of which true statements are true, the intermediate position would be that although the world is independent of what anyone believes, the facts are not, since what facts there are depend upon our capacities to verify (and hence on what we would believe under specifiable circumstances). This is an interesting idea. Ultimately I think the intermediate position is unstable, but to show that would be too substantial a task for the present context.

References

Austin, J. L. 1950. "Truth." *Proceedings of the Aristotelian Society*, suppl. vol. 24: 111–128. Reprinted in Pitcher, 1964, pp. 18–31.

Blackburn, S. W. 1984. *Spreading the Word*. Oxford: Clarendon Press.

Bradley, F. H. 1893. *Appearance and Reality*. London: Swan Sonnenschein.

Davidson, D. 1986. "A coherence theory of truth and knowledge." In *Truth and Interpretation*, edited by E. Lepore, pp. 307–319. Oxford: Blackwell. Reprinted in *Reading Rorty*, edited by A. Malachowski, pp. 120–134. Oxford: Blackwell, 1990.

Davidson, D. 1990a. "Afterthoughts, 1987." In *Reading Rorty*, edited by A. Malachowski, pp. 134–138. Oxford: Blackwell.

Davidson, D. 1990b. "The structure and content of truth." *Journal of Philosophy*, 87(6): 279–328.

Descartes, R. 1964–1976. "Meditation III." In *Oeuvres de Descartes*, rev. edn, edited by C. Adam and P. Tannery. Paris: Vrin/CNRS.

Dewey, J. 1938. *Logic: The Theory of Enquiry*. New York: Holt.

Dummett, M. A. E. 1978. *Truth and Other Enigmas*. London: Duckworth.

Frege, G. 1969. *Nachgelassene Schriften*, edited by H. Hermes, F. Kambartel, and F. Kaulbach. Hamburg: Felix Meiner. *Frege: Posthumous Writings*, translated by P. Long and R. White. Oxford: Blackwell, 1979.

Grover, D. 1992. *A Prosentential Theory of Truth*. Princeton: Princeton University Press.

Grover, D., J. Camp, and N. Belnap. 1975. "A prosentential theory of truth." *Philosophical Studies*, 27: 73–125. Reprinted in Grover, 1992, pp. 70–120.

Horwich, P. 1990. *Truth*. Oxford: Blackwell.

James, W. 1907. *Pragmatism*. New York: Longmans, Green.

Kant, I. 1781–1787. *Kritik der reinen Vemunft*. Riga: Hartknoch. *Critique of Pure Reason*, translated by N. Kemp Smith. London: Macmillan, 1929.

Kripke, S. 1975. "Outline of a theory of truth." *Journal of Philosophy*, 72(19): 690–716. Reprinted in *Recent Essays on Truth and the Liar Paradox*, edited by R. M. Martin, pp. 53–81. Oxford: Clarendon Press, 1984.

Mackie, J. L. 1973. *Truth, Probability and Paradox*. Oxford: Clarendon Press.

Misak, C. 1991. *Truth and the End of Enquiry*. Oxford: Clarendon Press.

Neurath, O. 1931. "Soziologie im Physikalismus." *Erkenntnis*, 2: 393–431. Translated as "Sociology and physicalism." In *Logical Positivism*, edited by A. J. Ayer, pp. 282–317. Glencoe, IL: Free Press, 1959. Also as "Sociology in the framework of physicalism." In *Neurath, Philosophical Papers 1913–1946*, edited by R. S. Cohen and M. Neurath, pp. 58–90. Dordrecht, Netherlands: Reidel, 1983.

Neurath, O. 1932–1933. *Protokollsatze, Erkenntnis*, 3: 204–214. Translated as "Protocol sentences." In *Logical Positivism*, edited by A. J. Ayer, pp. 199–208. Glencoe, IL: Free Press, 1959. Also as "Protocol statements." In *Neurath, Philosophical Papers 1913–1946*, edited by R. S. Cohen and M. Neurath, pp. 91–99. Dordrecht, Netherlands: Reidel, 1983.

Peirce, C. S. 1878. "How to make our ideas clear." *Popular Science Monthly*, 12: 286–302. Reprinted in his *Collected Papers*, vol. 5, pp. 248–271. Cambridge, MA: Belknap Press, 1931–1935.

Pitcher, G., ed. 1964. *Truth*. Englewood Cliffs, NJ: Prentice-Hall.

Prior, A. N. 1971. *Objects of Thought*. Oxford: Clarendon Press.

Putnam, H. 1978. "Realism and reason." In his *Meaning and the Moral Sciences*, pp. 123–140. London: Routledge and Kegan Paul.

Putnam, H. 1983. *Realism and Reason*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramsey, F. P. 1927. "Facts and propositions." *Proceedings of the Aristotelian Society*, suppl. vol. 7: 153–170. Reprinted in his *The Foundations of Mathematics*, 138–155. London: Kegan Paul, Trench, Trubner, 1931.

Russell, B. 1906–1907. "On the nature of truth." *Proceedings of the Aristotelian Society*, 7(1): 28–49.

Russell, B. 1956. "The philosophy of logical atomism." In his *Logic and Knowledge*, edited by R. C. Marsh, pp. 177–281. London: Allen and Unwin.

Strawson, P. F. 1950. "Truth." *Proceedings of the Aristotelian Society*, suppl. vol. 24: 129–156. Reprinted in his *Logico-Linguistic Papers*, pp. 190–213. London: Methuen, 1971. Also in Pitcher, 1964, pp. 32–43.

Strawson, P. F. 1965. "Truth: a reconsideration of Austin's views." *Philosophical Quarterly*, 15: 289–301. Reprinted in his *Logico-Linguistic Papers*, pp. 234–249. London: Methuen, 1971.

Tarski, A. 1935. "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica*, 1: 261–405. Reprinted as "The concept of truth in formalized languages." In his *Logic, Semantics, Metamathematics*, translated by J. H. Woodger, pp. 152–278. Oxford: Oxford University Press, 1956.
Tarski, A. 1944. "The semantic conception of truth." *Philosophy and Phenomenological Research*, 4: 341–375. Reprinted in *Readings in Philosophical Analysis*, edited by H. Feigl and W. Sellars, pp. 52–84. New York: Appleton-Century-Crofts, 1947. Also in *Semantics and the Philosophy of Language*, edited by L. Linsky, pp. 13–47. Champaign, IL: University of Illinois Press, 1952.
Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. London: Kegan Paul, Trench, Trubner.
Wright, C. 1987. *Realism, Meaning and Truth*. Oxford: Blackwell.
Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.
Wright, C. 1995. "Ralph C. S. Walker, the coherence theory of truth." *Synthese*, 103(2): 279–302.

Further Reading

Alston, W. P. 1996. *A Realist Conception of Truth*. Ithaca, NY: Cornell University Press.
Baldwin, T. 1991. "The identity theory of truth." *Mind*, 100(397): 35–52.
Blackburn, S. W. 1993. *Essays in Quasi-Realism*. New York: Oxford University Press.
Bradley, F. H. 1914. *Essays on Truth and Reality*. Oxford: Clarendon Press.
David, M. 1994. *Correspondence and Disquotation*. New York: Oxford University Press.
Devitt, M. 1984. *Realism and Truth*. Oxford: Blackwell.
Ellis, B. 1990. *Truth and Objectivity*. Oxford: Blackwell.
Field, H. 1972. "Tarski's theory of truth." *Journal of Philosophy*, 69(13): 347–375.
Field, H. 1988. "The deflationary conception of truth." In *Fact and Morality*, edited by G. MacDonald and C. Wright, pp. 55–117. Oxford: Blackwell.
Loar, B. 1980. "Ramsey's theory of belief and truth." In *Prospects for Pragmatism*, edited D. H. Mellor, pp. 49–70. Cambridge: Cambridge University Press.
Walker, R. 1989. *The Coherence Theory of Truth*. London: Routledge.
Walker, R. 1995. "Verificationism, anti-realism and idealism." *European Journal of Philosophy*, 3(3): 257–272.
Wright, C. 1988. "Realism, antirealism, irrealism, quasi-realism." *Midwest Studies in Philosophy*, 12(1): 25–49.

Postscript: Pluralism about Truth

MICHAEL P. LYNCH

One lesson of the above is that we seem to be able to say very little about truth in general. On the one hand, traditional theories seemed to overreach. They try to *reduce* truth to some other property, such as correspondence or coherence; as a consequence, they face counter-examples. On the other hand, broadly deflationary views (such as the redundancy theory, or the truistic version of the correspondence theory) just give up the game and admit that there is nothing informative to say about the nature of truth at all. Put starkly, the choice seems to be between two unattractive options: either truth has a single nature we can't define, or it is has no nature at all.

Over the past two decades, a third alternative has emerged. Keeping to the stark way of putting the matter, this third alternative takes it that truth has, in some sense, more than one nature, or that there is more than one way in which beliefs or propositions can be true.

This type of position has become known as *pluralism about truth*. It was originally proposed by Wright (1992; 1998; 2013) and then later developed, in a somewhat different way, by Lynch (2001a; 2004; 2009). Pluralist views typically involve two components: a non-reductive analysis of the concept of truth and an account of how more than one property can satisfy that concept.

The non-reductive account of the concept of truth favored by pluralists can be understood as a type of functionalism about the concept, according to which the secret to what truth is lies with what it does. Traditional theories of truth tended to celebrate certain intuitive features we take truth to have. The correspondence theory, for example, celebrates what we might call the Objectivity Principle: true beliefs "portray things as they are"; while the pragmatist theories celebrate the End of Inquiry Principle: true beliefs are what we aim to have at the end of inquiry. The pluralist suggests that we understand these principles not as revealing the single essence of truth, but as revealing an aspect of truth's *functional role*. Roughly speaking, "portraying things as they are" and "being the end of inquiry" are part of the truth-role. They, and other truisms about truth (such as the idea that it is correct to believe a proposition when it is true), reveal what true beliefs do. In short, the thought is that these principles tell us that true propositions are those that have a property that has a certain function in our cognitive economy.

Playing a functional role amounts to satisfying a description, one that picks out certain features possessed by anything that plays the role. Such descriptions are like job descriptions. Writing a job description involves listing the tasks anyone who has that job must do, and specifying how that job relates to others in the immediate economic vicinity. We define the job in terms of its place in a larger network of jobs, all of which are understood in relation to each other, and by weighting some aspects of the job as more important or crucial than others. In the philosophy of mind, where functionalist analyses are commonplace, so-called analytic functionalists take job descriptions for mental properties to be given by our implicit folk beliefs about those properties. In the case of a property like pain, these include truisms like "the threat of pain causes fear" and "if you are in pain, you may say 'ouch'" and "if you are hit in the head, you will probably be in pain" and so on. These platitudes tell us that a property plays the pain-role when it is related to certain other mental, behavioral, and experiential properties of an organism.[1]

Likewise, we can take the truth-role to be carved out by our common truths about truth. These truisms form a theoretical structure of sorts – one which illustrates the relationships between true propositions and propositions with various other properties such as warrant, belief, correctness, and so on. These features, as in the parallel case of functional properties in the philosophy of mind, will not be primarily causal in nature, but quasi-logical and explanatory. But the basic suggestion in both cases is the same: the unique relations that truth bears to other properties nonetheless suffice to pin it down by jointly specifying the truth-role.

This allows us to give truth-conditions for the application of the truth concept itself as follows:

(F) For every x, x is true if, and only if, x has a property that plays the truth-role.

Moreover, the functionalist can implicitly define the truth-role itself in terms of those relational features – call them the *truish features* – picked out by our common truths about

truth. Thus they can say, for example, that where a proposition P is T, T plays the truth-role just when: where P is believed, things are as they are believed to be; other things being equal, it is a worthy goal of inquiry to believe P if P is T; it is correct to believe P if and only if P is T.

One of the chief benefits of a functionalist analysis of the concept of truth is that it opens up a new way of understanding the metaphysics of truth. We saw above that properties like correspondence (understood in a robust way) aren't plausible candidates for the essence of truth. But they may well be excellent candidates to *play the truth-role*. Perhaps something like correspondence plays the truth-role for our beliefs about physical objects and their properties; perhaps something like superwarrant plays it for our normative beliefs.[2] In short, functionalism makes room for a pluralist metaphysics of truth, or the idea that different kinds of beliefs might be true in different ways.

A pluralist metaphysics of truth has significant implications. One such implication is that it would make room for the anti-deflationary idea that an appeal to truth – and its realizing properties – may be explanatorily useful. One obvious example is the nature of knowledge. If there is more than one property that can play the truth-role, then this fact might help to explain how and why moral knowledge differs in kind from knowledge about the things like cats and cars.

The functionalist view just described is more a theoretical framework than a full-fledged theory of truth. Individual advocates of the general approach can and do differ over how to fill in the details.

One thing they can differ over is the content of the common truths or platitudes that define the truth-role. Crispin Wright, an early advocate of this general approach, for example, has argued that equally or more fundamental than the truisms we've so far canvassed are platitudes linking truth with assertion and negation (Wright, 1992). Others have suggested that the number of platitudes required to demarcate the role is smaller, not larger, suggesting a view closer in spirit to some deflationary accounts.

Another area of significant disagreement concerns what to say about the nature of the truth property itself. Wright's work, for example, can be read as implying that truth is whatever property happens to play the truth-role (Wright, 1998). Thus, where that role is played by correspondence, truth is correspondence, where it is played by coherence, truth is coherence. But such a view seems to face significant problems. One such problem concerns its implication for our understanding of validity (Tappolet, 1997). According to the standard definition, valid inferences preserve a single property – truth. But now consider an inference like:

> Murder is wrong or two and two is five; two and two is not five.
> Therefore, murder is wrong.

Suppose (a) that "truth" denotes any given property that plays the truth-role and (b) that the truth-role is played by different properties in the moral and mathematical domains. If so, then the above argument does not preserve a single property contra our ordinary definition of validity.[3]

So it seems that we should not identify truth as such with the properties that realize the truth-role. A second suggestion (Pedersen and Wright, 2012) is that truth as such is a disjunctive property – a belief is true when it has the property of either corresponding or cohering. A third alternative (Lynch, 2001a) is to take "truth" to denote a higher-order

property – the property of having a property that plays the truth-role. This suggestion avoids the problem with mixed inferences – since "truth" would now denote a single property in all domains. But it comes at a price. For the point of the functionalist analysis was to define truth by way of a description of its functional role. That description implies that truth itself is a property that has the features described by our common truths about truth. But the property of having a property that has those features does not itself seem to have those features, contra our original analysis.[4]

One way out of this difficulty would be to say that truth just is the property that has the truish features essentially, but to allow that this property is itself *immanent* in its realizing properties (Lynch, 2009). Let us say that a property F is immanent in a property M if and only if it is *a priori* that F's conceptually essential features are a subset of M's features. Since it is *a priori* that every property's conceptually essential features are a subset of its own features, every property is immanent in itself. So immanence, like identity, is reflexive and transitive. But unlike identity, it is non-symmetric. Where M and F are ontologically *distinct* properties – individuated by non-identical sets of essential features and relations – and F is immanent in M, M is not thereby immanent in F.

There is more to say about immanence naturally, but even so briefly described, it seems like it might allow the functionalist what she wants.[5] Should a property of some belief such as *corresponding to reality* manifest truth, it will be *a priori* that the truish features are a subset of that property's features. Roughly put: corresponding to reality is what makes some beliefs true because being true is just part of what it is for some beliefs to correspond to reality.

The topic of truth is of perennial interest to philosophers, and involves more issues than we've had space to touch on in this chapter. Those include the liar paradox (see, e.g., Scharp, 2013), the ultimate tenability of deflationism (the articles in Lynch, 2001b), and the question of the value of truth (Lynch, 2004). Moreover, it remains an open question whether pluralism about truth in any form will prove to be successful. But it is promising. By leaving room for pluralism about the properties that play the truth-role, it avoids the scope problem plaguing traditional theories of truth. But by allowing that there is more to say about truth and its realizers than the deflationist allows, it suggests that understanding the metaphysics of truth can shed light on other issues of philosophical interest, such as the nature of knowledge or meaning. Suppose, for example, that the meaning of a sentence consists (at least in part) in its truth-conditions. If there is more than one kind of truth, then we would seem to have a clear explanation for the longstanding intuition that the meaning of ethical sentences differs in kind from the meaning of sentences about the physical world.[6]

Notes

1 For an overview of the functionalist options in the philosophy of mind, see, e.g., Kim, 1998, p. 146.
2 A belief is superwarranted, or what Crispin Wright calls superassertible, just when, roughly, its warrant would survive defeat. See Wright (1992) and Lynch (2009).
3 A related problem concerns "mixed compounds"; for discussion see Pedersen (2006), Pedersen and Wright (2012), Cotnoir (2013).
4 For other criticisms, see Wright (2010).
5 For a related view that has some of the same benefits, see Edwards (2013).
6 Thanks to Nathan Kellen for helpful comments.

References

Cotnoir, A. J. 2013. "Validity for strong pluralists." *Philosophy and Phenomenological Research*, 86(3): 563–579.

Edwards, D. 2013. "Truth as a substantive property." *Australasian Journal of Philosophy*, 91(2): 279–294.

Kim, J. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Lynch, M. P. 2001a. "A functionalist theory of truth." In Lynch, 2001b, pp. 723–749.

Lynch, M. P. 2001b, ed. *The Nature of Truth: Classic and Contemporary Perspectives*. Cambridge, MA: MIT Press.

Lynch, M. P. 2004. "Truth and multiple realizability." *Australasian Journal of Philosophy*, 82(3): 384–408.

Lynch, M. P. 2009. *Truth as One and Many*. Oxford: Oxford University Press.

Pedersen, N. J. L. L. 2006. "What can the problem of mixed inferences teach us about alethic pluralism?" *The Monist*, 89(1): 102–117.

Pedersen, N. J. L. L., and C. D. Wright. 2012. "Pluralist theories of truth." In *Stanford Encyclopedia of Philosophy*, edited by Kevin Scharp. https://plato.stanford.edu/entries/truth-pluralist/ (accessed December 7, 2016).

Pedersen, N. J. L. L., and C. D. Wright, eds. 2013. *Truth and Pluralism: Current Debates*. Oxford: Oxford University Press.

Scharp, K. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Tappolet, C. 1997. "Mixed inferences: a problem for pluralism about truth predicates." *Analysis*, 57(3): 209–210.

Wright, C. D. 2010. "Truth, Ramsification, and the pluralist's revenge." *Australasian Journal of Philosophy*, 88(2): 265–283.

Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Wright, C. 1998. "Truth: a traditional debate reviewed." *Canadian Journal of Philosophy*, 28: 31–74.

Wright, C. 2013. "A plurality of pluralisms." In Pedersen and Wright, 2013, pp. 123–153.

Further Reading

Bar-On, D., C. Horisk, and W. G. Lycan. 2000. "Deflationism, meaning and truth-conditions." *Philosophical Studies*, 101(1): 1–28.

Greenough, P., and M. Lynch. 2006. *Truth and Realism*. Oxford: Oxford University Press.

Hill, C. S. 2002. "Thought and world." In *An Austere Portrayal of Truth, Reference, and Semantic Correspondence*. Cambridge: Cambridge University Press.

Horwich, P. 1998. *Truth*. Oxford: Oxford University Press.

Künne, W. 2003. *Conceptions of Truth*. Oxford: Oxford University Press.

Lynch, M. P. 1998. *Truth in Context: An Essay on Pluralism and Objectivity*. Cambridge, MA: MIT Press.

22

Truthmaker Semantics

KIT FINE

My aim in the present chapter is to explain the basic framework of truthmaker or 'exact' semantics, an approach to semantics that has recently received a growing amount of interest, and then to discuss a number of different applications within philosophy and linguistics.

I Theory

1 Truthmakers in Metaphysics and Semantics

The idea of truthmaking is the idea of something on the side of the world – a fact, perhaps, or a state of affairs – verifying, or making true, something on the side of language or thought – a statement, perhaps, or a proposition. The idea of truthmaking has figured prominently in contemporary metaphysics and semantics. In its application to metaphysics, the thought has been that we can arrive at a satisfactory metaphysical view by attempting to ascertain *what it is*, on the side of the world, that renders true what we take to be true (as in Armstrong 1997 and 2004); and, on the semantical side, the thought has been that we can attain a satisfactory semantics for a given language by attempting to ascertain *how it is* that the sentences of the language are made true by what is in the world. In the former case, truthmaking serves as a conduit taking us from language or thought to an understanding of the world; and in the latter case, it has served as a conduit taking us from the world to an understanding of language.

I, along with other philosophers, have argued against truthmaking as a guide to metaphysics (Schnieder, 2006; Horwich, 2008; Fine, 2012a, §3). I have sometimes joked that truthmaking is fine as a guide to metaphysics as long as we junk the relata on the left, the things whose existence *makes* true, the relata on the right, the things *made* true, and the relation of *making* true. But my concern here is with truthmaking as a tool of semantics; and

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

it is worth remarking, in this regard, that the task of discerning truthmakers may be helpful for the one project even if not at all helpful for the other.

Indeed, the general focus of the two projects is very different. If our aim is to understand the world, then our focus should be on the ultimate truthmakers, on *what* in the world ultimately makes something true, and the question of *how* the truthmakers make the statements of our language true is of no great concern. But if our aim is to understand language, then our focus should be on the immediate truthmakers, not the ultimate truthmakers, and the question of *how* they make the statements of the language true will be of greatest concern. Take the statement 'there is a chair over there,' by way of example. For the metaphysical project, we may wish to give an account of the truthmakers for the statement in terms of elementary particles, let us say, and the real question of concern will be whether we can achieve a 'reduction' of the macroscopic to the microscopic. But for the semantical project, we can rest content with specifying the truthmakers in terms of ordinary macroscopic objects; and our concern in this case will not be with a reduction of the macroscopic to the microscopic but with what it is about the representational features of the statement itself that enables it to have the superficial truthmakers that it does. We see ripples on the surface of a pond; and our concern may be with what it is beneath the surface that causes the ripples or with how it is that the ripples play out over the surface, without regard for their cause.

2 *Truthmaker and Truth-Conditional Semantics*

There is a long tradition within philosophy, perhaps going all the way back to Frege (1892), of identifying the meaning of a statement with its truth-conditions, that is, with the conditions under which it is true. However, a truth-conditional account of meaning can take various different forms and it may be worthwhile to locate truthmaker semantics within a general account of this sort.

One major line of division concerns the form of the truth-conditional claims themselves. On the clausal approach, especially associated with Davidson (1967), the truth-conditions of a statement are not given as entities but by the clauses through which a theory of truth specifies when a statement is true. Thus a typical clause within such a theory might state that a conjunction 'A $\wedge$ B' is true when both of its conjuncts A and B are true. On the objectual approach, by contrast, the truth-conditions are objects, rather than clauses, which stand in a relation of truthmaking to the statements they make true. Within such an account, therefore, there is both an ontology of truthmakers and a relation of truthmaking.

Within the objectual approach, a second major line of division concerns the nature of the truthmakers. Under the most familiar version of the objectual approach, the truth-conditions of a statement are taken to be possible worlds and the content of a statement may, accordingly, be identified with the set of possible worlds in which it is true. This gives rise to 'possible worlds semantics,' which received its first systematic application to natural language in the work of Montague (1970).

Under a somewhat less familiar version of the objectual approach, the truth-conditions are not – or not, in general – taken to be possible worlds but states or situations – fact-like entities that serve to make up a world rather than being worlds themselves; and the content of a statement may, in this case, be identified with the set of verifying states or situations in which it is true. This gives rise to 'situation semantics,' which received its first systematic development in the work of Barwise and Perry (1983).

The main difference between the two kinds of semantics turns on the question of completeness: the truth-value of any statement will be settled by a possible world (at least to the extent that it is capable of being settled), whereas the truth-value of a statement may not be settled by a state or situation. The state of the weather in New York, for example, will not settle whether it is raining in London.

Within situation semantics itself, there is a third major line of division. For whereas it is tolerably clear what it takes for a statement to be true at a possible world, considerable unclarity surrounds the question of what it is for a statement to be true in a situation. There are at least three different conceptions of the truthmaking relation that one might adopt. We might call them *exact*, *inexact*, and *loose*; and they are successively broader. Thus each exact verifier of a statement is an inexact verifier and each inexact verifier a loose verifier.

Loose verification is a purely modal notion. A state or situation s will loosely verify a statement just in case the state necessitates the statement, that is, just in case it is impossible that the state obtain and the statement not be true. Exact and inexact verification, by contrast, require that there be a relevant connection between state and statement. With inexact verification, the state should at least be *partially* relevant to the statement; and with exact verification, it should be *wholly* relevant. Thus the presence of rain will be an exact verifier for the statement 'it is rainy'; the presence of wind and rain will be an inexact verifier for the statement 'it is rainy,' though not an exact verifier; and the presence of wind will be a loose verifier for the statement 'it is rainy or not rainy' (since the statement is true no matter what), while failing to be an inexact verifier. Loose and inexact verification are monotonic or 'hereditary': if a state necessitates or is partially relevant to the truth of a statement, then so is any more comprehensive state. But exact verification is not hereditary; the statement 'it is rainy' will be exactly verified by the presence of rain, for example, by not by the presence of rain and wind.

Here, in summary form, is a diagram of the various options:

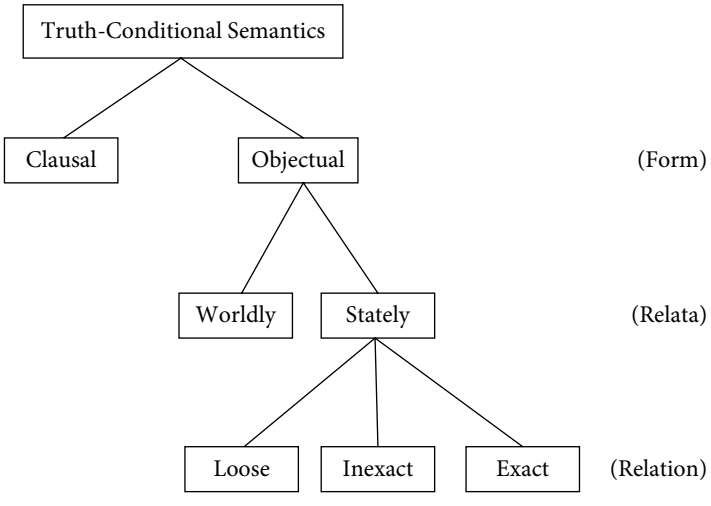

Figure 22.1

And within this framework, what I am calling 'truthmaker semantics' lies to the far right, with verification objectual, stately, and exact.

3 *Exact Verification*

When we survey the literature on situation semantics, we find that there has been little interest in the loose version of the semantics (though the work of Humberstone, 1981, and Rumfitt, 2012, is something of an exception). Most of the work – beginning with Barwise and Perry (1983) and extending to more recent work in the formal semantics for natural language (as typified by Kratzer, 2014) – has adopted the inexact approach or some variant of it.

The exact approach, on the other hand, has had a somewhat checkered career. Van Fraassen's "Facts and tautological entailments" (1969) was the first paper to present an exact semantics for classical logic. But it has been the fate of the semantics presented in that paper to be constantly rediscovered. Thus one reincarnation has been in the work of L. K. Schubert and his collaborators (Hwang and Schubert, 1993); another in the work of Stelzner (1992) on deontic logic; and another in some of the work of the school of Inquisitive Semantics at Amsterdam (Groenendijk and Roelofsen, 2010). I myself hit upon the idea of the semantics in the late 1960s and briefly mentioned its application to the problem of 'Disjunctive Simplification' in my review of Lewis's book on counterfactuals (Fine, 1975). However, it is only very recently that the full range and potential of the approach – as evidenced by the work of Aloni and Ciardelli (2013), Angell (1977; 1989; 2002), Fine (2012b; 2012c; 2013; 2014a; 2014b; 2015a; 2015b; 2015c; 2015d; 2015e; 2016), Gemes (1994; 1997), Moltmann (2007), van Rooij (2013), and Yablo (2014), among others – has begun to be appreciated.

It should also be mentioned, though this may not be evident to those working within only one of the respective fields, that there is an intimate connection between the recent semantical work on truthmaking and the recent metaphysical work on ground (Correia and Schnieder, 2012, is a recent collection of papers on ground). For if some statements A_1, A_2, … are to constitute a ground for another statement C, then they must be wholly relevant to the other statement's being the case; and, indeed, we might think of the notion of exact verification as being obtained through a process of ontological and semantic ascent from a claim of ground. For we first convert the statements A_1, A_2, … into the corresponding facts f_1, f_2, … (that A_1, A_2, …, obtain) and then take the sum f of the facts f_1, f_2, … to be an exact verifier for the truth of C. It is my belief that it is the notion of ground rather than the notion of truthmaking that is most relevant to the metaphysical project of discerning the nature of reality; and so it is of interest that this notion, of fundamental significance to metaphysics, can be 'reworked' in this way into a notion of fundamental significance to semantics.

4 *State Spaces*

In setting up the possible worlds semantics, we suppose given a 'pluriverse' of possible worlds; and, in applying the semantics, there is no need to suppose that possible worlds have any internal mereological structure – they can simply be taken to be undifferentiated 'blobs,' perhaps externally related by certain accessibility relations but with no internal mereological structure of their own.

For the purposes of setting up the truthmaker semantics, we suppose given a 'state space' of states. But it will be important, in applying the semantics, to suppose that the states, in contrast to possible worlds, are endowed with mereological structure. We must allow, for example, that one state may be a part of another, as when the presence of rain is part of the presence of rain and the wind, or that two states may fuse to a single composite state, as when the presence of rain and the presence of wind fuse to the presence of rain and wind.

It is also important in applying the semantics to appreciate that the term 'state' is a mere term of art and need not be a state in any intuitive sense of the term. Thus facts or events or even ordinary individuals could, in principle, be taken to be states, as long as they are capable of being endowed with the relevant mereological structure and can be properly regarded as verifiers. We allow the states, whatever they might be, to be possible as well as actual (as with Gore winning the presidency in 2001); and we even allow impossible states (as with Gore both winning and losing the presidency in 2001). Thus verification will have a counterfactual flavor; a verifying state is one that *would* make a given statement true were it to obtain, not necessarily one that *does* make the statement true.

From a purely mathematical point of view, we may take a *state space* to be an ordered pair $(S, \sqsubseteq)$, where S (states) is a non-empty set and $\sqsubseteq$ (part) is a binary relation on S. We suppose that the relation $\sqsubseteq$ is a *partial order*, that is, that it conforms to the following three conditions for any states s, t, and u of S:

(i) Reflexivity: $s \sqsubseteq s$
(ii) Anti-symmetry: $s \sqsubseteq t$ and $t \sqsubseteq s$ implies $s = t$
(iii) Transitivity: $s \sqsubseteq t$ and $t \sqsubseteq u$ implies $s \sqsubseteq u$.

We wish to impose a further condition on a state space, although this will call for two additional definitions. Given a subset $T \subseteq S$ of states, we say that s is an *upper bound of* T if it contains each state of T, that is, if $t \sqsubseteq s$ for each $t \in T$, and we say that s is a *least upper bound* (lub) *of* T if s is an upper bound of T and if it is included in any upper bound of T, that is, if $s \sqsubseteq s'$ for any upper bound s' of T. We then require that a state space be *complete* in the sense that every subset $T \subseteq S$ of states have a least upper bound.

The least upper bound of $T \subseteq S$ is unique (since if s and s' are least upper bounds, then $s \sqsubseteq s'$ and $s' \sqsubseteq s$ and so, by anti-symmetry, $s = s'$). We denote it by $\sqcup T$ and call it the *fusion of* T (or of the members of T). When $T = \{t_1, t_2, \ldots\}$, we may write $\sqcup T$ more perspicuously in the form $t_1 \sqcup t_2 \sqcup \ldots$; and so, in particular, $s \sqcup t$ will be the least upper bound of $\{s, t\}$. For many applications, we need only assume the existence of $s \sqcup t$ and not the existence of $t_1 \sqcup t_2 \sqcup \ldots$ for arbitrary $t_1, t_2, \ldots$.

The state space $(S, \sqsubseteq)$, as we have defined it, does not embody the distinction between possible and impossible states; for all that we have said, each state in S might be an impossible state. In order to give recognition to the distinction, we may take a *modalized state space* to be an ordered triple $(S, S^\Diamond, \sqsubseteq)$, where $(S, \sqsubseteq)$ is a state space as before and $S^\Diamond$ (possible states) is a non-empty subset of S. We make one assumption about $S^\Diamond$, namely:

Closure under Part $t \in S^\Diamond$ whenever $s \in S^\Diamond$ and $t \sqsubseteq s$.

Parts of possible states are also possible states. The interest of modalized state spaces is that they enjoy both a mereological structure (as represented by $\sqsubseteq$) and a modal structure (as represented by $S^\Diamond$); and the interaction of the two enables us to define many notions of significance. We may say, in particular, that two states s and t are *compatible* if their fusion $s \sqcup t$ is a possible state (i.e., a member of $S^\Diamond$) and *incompatible* if their fusion $s \sqcup t$ is not a possible state. Thus the possible state of my being cold and the possible state of my being hungry will presumably be compatible since the state of my being cold and hungry is a possible state, while the possible state of my being cold and the possible state of my being hot will presumably be incompatible since the state of my being both cold and hot is not a possible state.

We may also employ the resources of a modalized state space to say when a state corresponds to a possible world. For given a modalized space $(S, S^{\Diamond}, \sqsubseteq)$, we may say that the state s is a *world-state* if it is possible and if any state is either a part of s or incompatible with s. Thus a world-state must positively include or exclude any other state. Given the notion of a world-state, we may then say that the space $(S, S^{\Diamond}, \sqsubseteq)$ is a *W-space* if every possible state of S is part of a world-state. Thus a W-space will, in effect, contain the pluriverse of possible worlds. However, very few applications require the assumption that the state space be a W-space and so, from this perspective, the postulation of possible worlds is a gratuitous assumption that serves no real purpose.

It should be noted that our approach to states is highly general and abstract. We have formed no particular conception of what they are; and nor have we assumed that there are 'atomic' states, from which all other states can be obtained by fusion. Nearly all of the existing literature on the topic has failed to adopt such a neutral perspective. Thus states are often identified with sets of possible worlds (where the worlds themselves might be identified with sets of sentences) or it is assumed that all states are constructed from atomic states which are somehow isomorphic with the atomic sentences of the language under consideration.

Nothing is gained by this lack of generality or abstraction and a great deal is lost. For one thing, the particular identifications or assumptions may not be appropriate in certain contexts. One might well think, for example, that a progressive statement such as 'this is moving' is not made true by any atomic state but by the motions of the object over successively shorter intervals of time; and if one takes a state to be a set of possible worlds, then one denies oneself the possibility of distinguishing between different necessary states or different impossible states. The technical development of the subject also requires a more abstract approach. For one will want to perform certain constructions on state spaces (forming product spaces, for example, or congruent spaces) in which the special identifications of or restrictions on the original spaces are lost.

The abstract approach to modal logic championed by Kripke's early work (in which possible worlds are simply regarded as arbitrary points, rather than as models or sets of sentences) has been a great boon to the formal development of modal logic; and it is to be hoped that future researchers will appreciate that there are similar benefits to be gained by adopting a more abstract approach to the truthmaker framework as well.

5 *Sentential Semantics for Exact Verification*

A good test for any proposed semantical framework is its ability to deal with classical sentential logic; and so let me show how we might give such a semantics within the truthmaker framework.

Recall the standard possible worlds semantics for sentential logic. For the case of negative, conjunctive, and disjunctive statements we have the following clauses:

(ii) a world verifies $\neg$A iff it does not verify A;
(iii) a world verifies a conjunction A $\wedge$ B iff it verifies A and verifies B;
(iv) a world verifies a disjunction A $\vee$ B iff it verifies A or verifies B.

Let us now give the corresponding clauses under the truthmaker approach. Our aim is not simply to say when a statement is true *at* a world but to say what it is *in* the world that

makes it true and in such a way that the truthmaker is wholly relevant to the statement it makes true. To this end, we shall find it helpful to give separate clauses for when a statement is verified and for when it is falsified.[1] Here are the clauses for negation, conjunction, and disjunction, which are split into two parts – one for verification and the other for falsification:

(ii)′ A state verifies a negative statement ¬A just in case it falsifies the negated statement A;
a state falsifies the negative statement ¬A just in case it verifies the negated statement A.

(iii)′ A state verifies a conjunction A ∧ B just in case it is the fusion of states that verify the respective conjuncts A and B;
a state falsifies the conjunction A and B just in case it falsifies A or falsifies B.

(iv)′ A state verifies a disjunction A ∨ B just in case it verifies one of the disjuncts A or B;
a state falsifies the disjunction A ∨ B just in case it is the fusion of states that falsify the respective disjuncts A and B.

Clearly, these clauses are all very plausible, once we have in mind that our interest is in the notion of *exact* verification: what exactly verifies a conjunction is the fusion of the verifiers for its conjuncts; what exactly verifies a disjunction is a verifier for one of the disjuncts; and similarly for the falsification of a conjunction or of a disjunction.

Let us now provide a more technical exposition of the semantics. We suppose given an infinitude of atomic sentences p_1, p_2, …. Formulas of the sentential language are then constructed from the atomic sentences using the Boolean operators ∧, ∨, and ¬ in the usual way.

A (*state*) *model* $\boldsymbol{M}$ is an ordered triple $(S, \sqsubseteq, |\cdot|)$, where $(S, \sqsubseteq)$ is a state space, as before, and $|\cdot|$ (valuation) is a function taking each sentence letter p into a pair (V, F) of subsets of S – intuitively, the set $|p|^+ = V$ of its verifiers and the set $|p|^- = F$ of its falsifiers.

When a model $\boldsymbol{M} = (S, S^\Diamond, \sqsubseteq, |\cdot|)$ is constructed over a modalized state space $\boldsymbol{S} = (S, S^\Diamond, \sqsubseteq)$, one might want to take into account the interaction between the verifiers V and the falsifiers F of any given atomic sentence p. There are then two plausible conditions that might be imposed:

Exclusivity: No verifier is compatible with a falsifier (i.e., no member of V is compatible with a member of F);
Exhaustivity: Any possible state is compatible with a verifier or with a falsifier (i.e., each possible state is compatible with a member of V or with a member of F).

Exclusivity corresponds to the assumption that no statement is both true and false (which is how things would be if a verifier were compatible with a falsifier); and Exhaustivity corresponds to the assumption that every statement is either true or false (since no possible state could exclude the statement being either true or false). Of course, either of these assumptions could be dropped if one wanted to allow either truth-value gluts (a statement being both true and false) or truth-value gaps (a statement being neither true nor false).

Given a model $\boldsymbol{M} = (S, \sqsubseteq, |\cdot|)$, we may now define what it is for an arbitrary formula A to be *verified by* a state s (s ||- A) or *falsified by* a state s (s -|| A):

(i)$^+$ s ||- p if $s \in |p|^+$;
(i)$^-$ s -|| p if $s \in |p|^-$;
(ii)$^+$ s ||- ¬B if s -|| B;
(ii)$^-$ s -|| ¬B if s ||- B;
(iii)$^+$ s ||- B ∧ C if for some states t and u, t ||- B, u ||- C, and $s = t \sqcup u$;
(iii)$^-$ s -|| B ∧ C if s -|| B or s -|| C;
(iv)$^+$ s ||- B ∨ C if s ||- B or s ||- C;
(iv)$^-$ s -|| B ∨ C if for some t and u, t -|| B, u -|| C, and $s = t \sqcup u$.

Here (ii)$^+$ and (ii)$^-$ correspond to (ii)′ above, (iii)$^+$ and (iii)$^-$ to (ii)′, and (iv)$^+$ and (iv)$^-$ to (iv)′.

6 *Some Features and Consequences of the Semantics*

I should like to discuss some noteworthy features of the semantics, especially in so far as they stand in contrast to those of other, more familiar, semantical schemes.

One immediate odd consequence of the clauses is that p ∧ p may not have the same verifiers as p. For suppose that s and t are the two verifiers of p, where neither is a part of the other. Then $s \sqcup t$ is a verifier for p ∧ p by clause (iii)$^+$ but not a verifier for p. I believe that this odd consequence is not a problem for natural language applications but that it may be a problem for more theoretical applications. If, for example, our interest is in the partial truth or confirmation or verisimilitude of a statement, then it seems to me that we would not want to distinguish in this way between the statements p and p ∧ p.

We may avoid this odd consequence, if we wish, by requiring that a verifier for A ∧ B should also be a verifier for A ∨ B (and, likewise, by requiring that a falsifier of A ∨ B should also be a falsifier for A ∧ B). Thus (iii)$^-$ and (iv)$^+$ now become:

(iii)$^{*-}$ s -|| B ∧ C if s -|| B or s -|| C or s -|| B ∨ C
(iv)$^{*+}$ s ||- B ∨ C if s ||- B or s ||- C or s ||- B ∧ C.

and we obtain what might be called the 'inclusive' semantics. It is characteristic of the inclusive semantics that the set $\{s \in S: s \text{ ||- A}\}$ of verifiers of any statement A should always be *closed under fusion*, that is, if $s_1, s_2, \ldots$ are one or more verifiers of A then so is their fusion $s_1 \sqcup s_2 \sqcup \ldots$.

There are a number of other variants of the clauses that might also be considered (as in Fine, 2015b, or Groenendijk and Roelofsen, 2010) but, in the interests of simplicity, we shall usually confine our attention in what follows to the original non-inclusive clauses.

It is evident from the above account that exact verification need not be hereditary. Indeed, there is nothing to prevent an atomic sentence p being exactly verified by a single state p and yet not verified by any distinct state $p' \sqsupseteq p$. However, if the verifiers of all the atomic sentences $p_1, p_2, \ldots$ are hereditary, that is, if s ||- p_k and $s' \sqsupseteq s$ implies s' ||- p_k, then it can be shown that the verifiers of all formulas A whatever will be hereditary, that is, that s ||- A and $s' \sqsupseteq s$ implies s' ||- A. Thus the failure of hereditariness arises from the behavior of the atomic sentences and is not attributable to the behavior of the sentential connectives.

It should also be evident that there is no reason in general to think that the exact verifiers will be minimal. Say that the state s *minimally verifies* the formula A if s exactly verifies A and if no proper part of s exactly verifies A (i.e., if $s' \sqsubseteq s$ and $s' \Vdash A$ implies $s' = s$). Now suppose that p is the sole verifier of p and q the sole verifier of q, with $q \not\sqsubseteq p$. Then p and $p \sqcup q$ are both verifiers of $p \vee (p \wedge q)$, with $p \sqcup q$ non-minimal since it contains the verifier p as a proper part. Indeed, there is no reason to suppose that a statement with verifiers need have any minimal verifiers at all. In the case of 'this is moving,' for example, we may well maintain that any verifier (the motion of the object through an interval of time) will contain another verifier as a proper part.

There has been a persistent tendency in the literature (we might call it 'minimalitis') to start off with a hereditary notion of verification and then attempt to get the corresponding notion of minimal verification, or some variant of it, to do the work of exact verification (as in the account of 'exemplification' in Kratzer, 2014). But if I am correct, all such attempts are doomed to failure. The relevant sense in which an exact verifier is wholly relevant to the statement it makes true is not one which requires that no part of the verifier be redundant but is one in which each part of the verifier can be seen to play an active role in verifying the statement. Thus the verifier $p \sqcup q$ of $p \vee (p \wedge q)$ can be seen to play such an active role, even though the part q is redundant, because of its connection with the second disjunct $(p \wedge q)$.

It is important to note that within the present semantics (and this is also true of a number of variants), two formulas A and B may have the same verifiers while $\neg A$ and $\neg B$ do not have the same verifiers. For let A be the formula $p \wedge (q \vee r)$ and B the formula $(p \wedge q) \vee (p \wedge r)$. Then it is readily verified that A and B will have the same verifiers (in any model). For the verifiers of $p \wedge (q \vee r)$ will be the fusions $p \sqcup s$ of a verifier p for p and a verifier s for $(q \vee r)$ by clause $(iii)^+$ above and hence will be the fusions $p \sqcup q$ of a verifier p for p and a verifier q for q or the fusions $p \sqcup r$ of a verifier p for p and a verifier r for r by clause $(iv)^+$, and these are exactly the verifiers of $(p \wedge q) \vee (p \wedge r)$ (again by $(iii)^+$ and $(iv)^+$). However, the verifiers for $\neg A$ and $\neg B$ may not be the same in this case. For suppose p is the sole falsifier of p, q of q, and r of r. Then by clauses $(iii)^-$ and $(iv)^-$, $p \sqcup r$ is a falsifier of $(p \wedge q) \vee (p \wedge r)$ but not of $p \wedge (q \vee r)$ (unless, of course, $p \sqcup r$ happens to be identical to p or to $q \sqcup r$).

This means that it is essential to define verification by means of a double induction on verification and falsification, as we did above, since the verifiers of the negation statement $\neg A$ cannot in general be determined from the verifiers of the negated statement A. It also means that we must complicate our definition of *proposition*. Within the possible worlds framework, the proposition (or content) expressed by a bivalent statement A can be identified with the set of worlds at which it is true. There is no need to bring in the worlds at which it is false, since these will simply be the worlds at which it is not true. Within the present framework, we might similarly identify the proposition expressed by a bivalent statement A with the set of its verifiers $V = \{s \in S: s \Vdash A\}$. We thereby obtain what I call a *unilateral* conception of propositionhood. But this conception will not be adequate if we wish to able to discern the proposition expressed by $\neg A$ from the proposition expressed by A. In this case, we should adopt a *bilateral* conception of propositionhood, according to which the proposition expressed by a statement A is a *pair* (V, F) of sets of states consisting of its set of verifiers $V = \{s \in S: s \Vdash A\}$ and its set of falsifiers $F = \{s \in S: s \dashv\!\Vdash A\}$ (in effect, this was already presupposed in our previous account of the valuation function $|\cdot|$). Under the unilateral conception of propositions, we might then take the conjunction $P \wedge Q$ of two propositions P and Q to be $\{p \sqcup q: p \in P \text{ and } q \in Q\}$ and we might take their disjunction $P \vee Q$

to be $P \cup Q$ (i.e., $\{s: s \in P \text{ or } s \in Q\}$); and, under the bilateral conception of propositions, we might take the negation of the proposition (P, P') to be (P', P), the conjunction of the propositions (P, P') and (Q, Q') to be $(P \wedge Q, P' \vee Q')$, and their disjunction to be $(P \vee Q, P' \wedge Q')$ (in conformity with clauses $(ii)^+$–$(iv)^-$ above).

One interesting aspect of the present approach is that it enables us to *define* the notions of inexact and loose verification. Thus we may say that *s inexactly verifies* A – in symbols, $s \,||{>}\, A$ – if *s* contains an exact verifier of A, that is, if for some state $s' \sqsubseteq s$, $s' \,||\text{-}\, A$, and we may say (within a modalized model) that *s loosely verifies* A – in symbols, $s \,|{=}\, A$ – if *s* is incompatible with any exact falsifier of A, that is, if *s* is incompatible with *t* whenever $t \text{ -}|| A$. It turns out that the 'reverse' definitions are not possible: exact verification cannot be defined in terms of inexact verification; and inexact verification cannot be defined in terms of loose verifications. Thus the different notions of verification – loose, inexact, and exact – involve a progressively greater commitment to what might be involved in the verification of a given statement; and we obtain the greatest flexibility in developing a theory of verification by taking the exact notion as primitive and seeing the other notions as off-shoots of the exact notion.

A further interesting aspect of the approach is that it naturally gives rise to two notions of consequence. For there are two natural ways of defining the relation of consequence between propositions *P* and *Q* – in terms either of conjunctive part or of disjunctive part. We may say that *P* is a *conjunctive part of Q* if *Q* is the proposition $P \wedge R$ for some proposition *R*; and we may say that *P* is a *disjunctive part of Q* if *Q* is the proposition $P \vee R$ for some proposition *R*. We may then define *Q* to be a consequence of *P* either when *Q* is a conjunctive part of *P* or when *P* is a disjunctive part of *Q*.

Under the identification of classical truth-functionally equivalent propositions (and hence, in particular, under the possible worlds approach), the two relations of consequence will coincide. For if *Q* is a conjunctive part of *P*, that is, if *P* is the proposition $Q \wedge R$ for some *R*, then *Q* will be the proposition $P \vee Q$ (given the classical truth-functional equivalence of *Q* to $(Q \wedge R) \vee Q$) and hence *P* will be a disjunctive part of *Q*; and if *P* is a disjunctive part of *Q*, that is, if *Q* is the proposition $P \vee R$, then *P* will be the proposition $Q \wedge P$ (given the classical truth-functional equivalence of *P* to $(P \vee R) \wedge P$) and hence *Q* will be a conjunctive part of *P*.

However, within the truthmaker approach the two relations come apart. Let us, for simplicity, work with a unilateral conception of propositions, under which a proposition is identified with its set of verifiers. Then the relation of *P*'s being a disjunctive part of *Q* will correspond most closely to the classical notion of consequence, since this is the relation that will hold between *P* and *Q* when every verifier of *P* is a verifier of *Q* (just as *Q* will be a consequence of *P* under the possible worlds approach when every world of *P* is a world of *Q*). But the notion of *P*'s being a conjunctive part of *Q* will be very different and has no real counterpart within the possible worlds approach. *P* will be a conjunctive part of *Q* only if (and, under certain simplifying assumptions, if and only if) the following two conditions obtain:

(i) every verifier of *P* is included in a verifier of *Q*; and
(ii) every verifier of *Q* contains a verifier of *P*.

Thus the proposition $P = \{p, q\}$ will be a conjunctive part of the proposition $Q = \{p \sqcup r, p \sqcup s, q \sqcup r, q \sqcup s\}$, since *Q* is the conjunction of $P \wedge R$, for *R* the proposition $\{r, s\}$. But, of course, *Q* is not, in general, a disjunctive part of *P*.

The relation of conjunctive part corresponds to the intuitive notion of *partial content* – of what is conveyed, in whole or part, by what is said. Thus in saying that Fido is a cocker spaniel, I convey, in part, that he is a spaniel and that he is a dog, but I do not convey, even in part, that he is a dog or a cat. In the one case, the content of 'Fido is a spaniel' is part of the content of 'Fido is a cocker spaniel' while, in the other case, the content of 'Fido is a dog or a cat' is not part of the content of 'Fido is a dog.'

The existence of the two relations of consequence may be of some methodological significance to the study of linguistics. For it is often assumed that intuitions of validity provide a key piece of data (some might think, *the* key piece of data) in the construction of a formal semantics for natural language. But, if I am right, then we should be somewhat more sensitive to the different inferential relationships that might be in play and it will be of particular importance to distinguish the subclass of inferential relationships that preserve content (as in the example above) and not merely truth.

There is one final aspect of truthmaker semantics which I should mention and which is of the greatest importance. It will have been noted that in specifying the verifiers of truth-functionally complex statements, we have not restricted ourselves to possible states. For given that p is a verifier for p and q is a verifier for q, we take $p \sqcup q$ to be a verifier for $p \wedge q$ even if p and q are incompatible and $p \sqcup q$ is therefore an impossible state. But it is not just that we do not restrict ourselves to possible states, we do not even make use of the distinction between possible and impossible states. The distinction is simply irrelevant in specifying the semantics for the connectives.

Of course, the distinction between possible and impossible will be required in providing an account of modal notions. We may want to say, for example, that 'necessarily, A' is true if every *possible* state is compatible with a verifier of A. Or again, we will say that s loosely verifies A (a modal notion) if s is *incompatible* with any falsifier of A and that C is a classical consequence of A (another modal notion) if every loose verifier of A is a loose verifier of C. But the present point of view is that there is nothing in the general notion of content or meaning or in the most general logical devices that requires us to draw the distinction between possible and impossible states. This freedom from the modal thinking that has been so characteristic of the more usual approaches to semantics is, I believe, one of the most distinctive and liberating aspects of the present approach.

7 Quantifiers

I have so far said nothing about the quantifiers. There are a number of different options for extending the clauses for the connectives to the quantifiers. But rather than considering them all, let me discuss one especially simple option, with an indication of how it might be extended to other cases.

One very general strategy for providing a semantics for quantificational statements is to reduce them to the corresponding truth-functional statements. Thus suppose the universal quantifier $\forall x$ ranges over the individuals $a_1, a_2, \ldots$.Then given that $a_1, a_2, \ldots$ are constants for the corresponding individuals $a_1, a_2, \ldots$, we might take the content of $\forall x \varphi(x)$ to be the same as the content of the conjunction $\varphi(a_1) \wedge \varphi(a_2) \wedge \ldots$; and, similarly, we might take the content of the existential quantification $\exists x \varphi(x)$ to be the same as that of the disjunction $\varphi(a_1) \vee \varphi(a_2) \vee \ldots$.

If we apply this general strategy to the present case, we arrive at the following clauses for the two quantifiers:

(v) a state verifies $\forall x\varphi(x)$ if it is the fusion of verifiers of its instances $\varphi(a_1)$, $\varphi(a_2)$, …;
a state falsifies $\forall x\varphi(x)$ if it falsifies one of its instances.
(vi) a state verifies $\exists x\varphi(x)$ if it verifies one of its instances $\varphi(a_1)$, $\varphi(a_2)$, …;
a state falsifies $\exists x\varphi(x)$ if it is the fusion of falsifiers of its instances.

Within a more formal treatment, we might introduce variables, predicates, and quantifiers into the language and define a formula (of the resulting first-order language) in the usual way. A *model* ***M*** is now an ordered quadruple $(S, A, \sqsubseteq, |\cdot|)$, where $(S, \sqsubseteq)$ is a state space, as before, A (individuals) is a non-empty set, and $|\cdot|$ (valuation) is a function taking each n-place predicate F and any n individuals $a_1, a_2, \ldots, a_n$ of A into a pair (V, F) of subsets of S – where, intuitively, V is the set of states which verifies F of $a_1, a_2, \ldots, a_n$ and F is the set of states which falsify F of $a_1, a_2, \ldots, a_n$.

We may now introduce constants $a_1, a_2, \ldots$ into the language, one for each of the distinct individuals $a_1, a_2, \ldots$ that compose A. We then have the following clauses for the closed atomic and quantificational formulas:

(i)$^+$ $s \Vdash Fa_1a_2 \ldots a_n$ if $s \in |F, a_1, a_2, \ldots, a_n|^+$;
(i)$^-$ $s \nVdash Fa_1a_2 \ldots a_n$ if $s \in |F, a_1, a_2, \ldots, a_n|^-$
(v)$^+$ $s \Vdash \forall x\varphi(x)$ if there are states $s_1, s_2, \ldots$ with $s_1 \Vdash \varphi(a_1), s_2 \Vdash \varphi(a_2), \ldots$ and $s = s_1 \sqcup s_2 \sqcup \ldots$;
(v)$^-$ $s \nVdash \forall x\varphi(x)$ if $s \nVdash \varphi(a)$ for some individual $a \in A$.
(vi)$^+$ $s \Vdash \exists x\varphi(x)$ if $s \Vdash \varphi(a)$ for some individual $a \in A$;
(vi)$^-$ $s \nVdash \exists x\varphi(x)$ if there are states $s_1, s_2, \ldots$ with $s_1 \nVdash \varphi(a_1), s_2 \nVdash \varphi(a_2), \ldots$ and $s = s_1 \sqcup s_2 \sqcup \ldots$.

Just as we can allow for a more inclusive clause for the falsification of a conjunction or the verification of a disjunction (clauses (iii)$^{*-}$ and (iv)$^{*+}$ from §I6 above), so we can allow for a more inclusive clause for the falsification of a universal quantification and the verification of an existential quantification:

(v)$^{*-}$ $s \nVdash \forall x\varphi(x)$ if for some distinct individuals $a_{k1}, a_{k2}, \ldots, s_{k1} \nVdash \varphi(a_{k1}), s_{k2} \nVdash \varphi(a_{k2}), \ldots$, and $s = s_{k1} \sqcup s_{k1} \sqcup \ldots$;
(vi)$^{*+}$ $s \Vdash \exists x\varphi(x)$ if for some distinct individuals $a_{k1}, a_{k2}, \ldots, s_{k1} \Vdash \varphi(a_{k1}), s_{k2} \Vdash \varphi(a_{k2}), \ldots$, and $s = s_{k1} \sqcup s_{k1} \sqcup \ldots$.

The difference between the two sets of clauses is that, in the first case, a universal quantification is only falsified and an existential quantification only verified via a single instance while, in the second case, the quantificational statements are verified or falsified via one or more instances.

One problem with these clauses is that they presuppose a fixed domain of individuals. For suppose that the actual individuals are $a_1, a_2, \ldots$ and that a is a merely possible individual (distinct from each of $a_1, a_2, \ldots$). Then in a possible world in which a exists, the truth of the instances $\varphi(a_1), \varphi(a_2), \ldots$ is not sufficient to guarantee the truth of $\forall x\varphi(x)$ and hence the fusion of verifiers for $\varphi(a_1), \varphi(a_2), \ldots$ need not be a verifier for $\forall x\varphi(x)$ (this is a familiar problem, going back to Russell, 1918/1919, pp. 236–237, and the early days of logical atomism).

There are a number of ways in which one might attempt to solve this problem within the present framework. One might drop the requirement that a verifying state should necessitate the statement that it verifies, for example, or one might treat the universal statement $\forall x\varphi(x)$ as equivalent, in effect, to $\Pi\, x(\neg Ex \vee \varphi(x))$, where $\Pi\, x$ is an 'outer' quantifier ranging over a fixed domain of all actual and possible individuals and E is an existence predicate – thereby reducing the variable domain case to the fixed domain case. My own favored solution is to suppose that, for each subdomain B of possible individuals, there is a totality state τ_B to the effect that the individuals of B are exactly the individuals that there are. The reader should note that the totality state τ_B is taken to exist even when it is not possible for the individuals of B to be exactly the individuals that there are.[2] The two clauses for the universal quantifier then take the following form:

s ||- $\forall x\varphi(x)$ if there is a subset $B \subseteq A$ composed of the distinct individuals $b_1, b_2, \ldots$ and states $s_1, s_2, \ldots$ such that s_1 ||- $\varphi(b_1)$, s_2 ||- $\varphi(b_2)$, … and $s = \tau_B \sqcup s_1 \sqcup s_2 \sqcup \ldots$;
s -|| $\forall x\varphi(x)$ if there is a subset $B \subseteq A$, an individual b from B, and a state s' such that s' -|| $\varphi(b)$ and $s = \tau_B \sqcup s'$.

And similarly for the existential quantifier.

Another problem is that we have only provided clauses for unrelativized quantification. But one might also want to deal with relativized quantifiers – as in $\forall x(\varphi(x)\colon \psi(x))$ (all φ's ψ) and $\exists x(\varphi(x)\colon \psi(x))$ (some φ's ψ). To take care of this problem, one might relativize the totality condition to the suitably defined content $|\varphi(x)|$ of the restrictor $\varphi(x)$. The clauses for the universal quantifier would then take the following form:

s ||-$\forall x(\varphi(x)\colon \psi(x))$ if there is a subset $B \subseteq A$ composed of the distinct individuals $b_1, b_2, \ldots$ and states $s_1, s_2, \ldots$ such that s_1 ||- $\psi(b_1)$, s_2 ||- $\psi(b_2)$, … and $s = \tau_{|\varphi(x)|,B} \sqcup s_1 \sqcup s_2 \sqcup \ldots$;
s -|| $\forall x(\varphi(x)\colon \psi(x))$ if there is a subset $B \subseteq A$, an individual b from B, and a state s' such that s' -|| $\psi(b)$ and $s = \tau_{|\varphi(x)|,B} \sqcup s'$.

We should note that, on this account, the verifiers for $\forall x(\varphi(x)\colon \psi(x))$ and for $\forall x(\neg\varphi(x) \vee \psi(x))$ will differ, since the first will involve a relativized totality condition $\tau_{|\varphi(x)|,B}$ to the effect that the individuals of B are exactly the individuals that φ, while the second will involve an unrelativized condition τ_B to the effect that the individuals of B are exactly the individuals that there are.

Two variants on the preceding accounts should be briefly noted. First, we have taken a verifier s of $\forall x\varphi(x)$, for example, to be the *fusion* of a totality condition τ_B and the verifiers $s_1, s_2, \ldots$ of the instances of $\forall x\varphi(x)$. But one might think of the totality condition as a precondition for the fusion $s_1 \sqcup s_2 \sqcup \ldots$ of $s_1, s_2, \ldots$ to verify $\forall x\varphi(x)$; and this is a reason for thinking of the verifier s of $\forall x\varphi(x)$, not as the fusion of τ_B and $s_1 \sqcup s_2 \sqcup \ldots$, but as an ordered pair $< \tau_B, s_1 \sqcup s_2 \sqcup \ldots >$, whose first component τ_B is a precondition and whose second component $s_1 \sqcup s_2 \sqcup \ldots$ is a post-condition or verifier proper. It turns out that this way of articulating a verifier into pre- and post-condition is very useful in a number of different contexts (and especially to presupposition).

Second, the verifiers of $\forall x\varphi(x)$ (or $\forall x(\varphi(x)\colon \psi(x))$) have been taken to involve the particular individuals $a_1, a_2, \ldots$ in the range of the quantifier. But it might be thought that $\forall x\varphi(x)$ is verified, in the first place, by certain general facts which, in themselves, do not involve any particular individuals. It turns out that this idea of generic verification can be

developed within the framework of arbitrary objects developed in Fine (1985). Thus the verifier of 'all men are mortal' might be taken to be the generic fact that the arbitrary man is mortal. Although I shall not go into the matter, I believe that the idea of generic verification has a wide range of interesting applications and is especially relevant to problems of 'logical omniscience.'

II Applications

I shall consider five applications in all: two to philosophical logic – the logic of partial content and subject-matter; and three to linguistics – counterfactuals, imperatives, and scalar implicature.

1 The Logic of Partial Content

We have already noted that there appears to be an intuitive sense in which the content of one statement is part of the content of another, for which Q will generally be part of the content of $P \wedge Q$ but $P \vee Q$ will not generally be part of the content of P. In a series of publications dating from 1977, Angell developed a system of 'analytic implication' that was intended to capture the logic of this notion (Angell, 1977; 1989; 2002). It is therefore natural to wonder what the relationship is between his system and our own semantical account of the notion.

It turns out that his system exactly corresponds to our own account, at least under certain very natural assumptions. Let us take $\neg$, $\wedge$, and $\vee$, as before, to be the primitive truth-functional connectives. The notion of the content of A containing the content of B ($A \rightarrow B$) and the notion of the content of A being equivalent to the content of B ($A \leftrightarrow B$) are interdefinable (with $(A \rightarrow B) =_{df} A \wedge B \leftrightarrow A$ and $(A \leftrightarrow B) =_{df} (A \rightarrow B) \wedge (B \rightarrow A)$); and let us, for convenience, take content equivalence ($\leftrightarrow$) rather than content containment $\rightarrow$ as primitive.

Angell's system may then be axiomatized by means of the following axioms and rules:

A1 $A \leftrightarrow \neg\neg A$
A2 $A \leftrightarrow A \wedge A$
A3 $A \wedge B \leftrightarrow B \wedge A$
A4 $(A \wedge B) \wedge C \leftrightarrow A \wedge (B \wedge C)$
A5 $\neg(A \wedge B) \leftrightarrow (\neg A \vee \neg B)$
A6 $\neg(A \vee B) \leftrightarrow (\neg A \wedge \neg B)$
A7 $A \wedge (B \vee C) \leftrightarrow (A \wedge B) \vee (A \wedge C)$
R1 $A(B), B \leftrightarrow C / A(C)$

The sole rule is R1, which allows us to substitute provable equivalents within any given theorem.

We now adopt the inclusive semantics given under §I.6 above and insist that, in any model $\boldsymbol{M} = (S, \sqsubseteq, (\cdot))$, the set of verifiers and the set of falsifiers of any atomic formula should be non-empty and closed under fusion (it can be shown that this requirement will be met by all formulas whatever if it is met by the atomic formulas).

Let us say that A *analytically implies* B in a model $\boldsymbol{M} = (S, \sqsubseteq, (\cdot))$ if (i) every exact verifier of A (in $\boldsymbol{M}$) contains an exact verifier of B and (ii) every exact verifier of B is contained in

an exact verifier of A (in conformity with our previous account of partial content). Now say that the formula A ↔ B *holds in* the model $\boldsymbol{M} = (S, \sqsubseteq, (\cdot))$ if A analytically implies B and B analytically implies A in $\boldsymbol{M}$. Finally, say that the formula A ↔ B is *valid* if it holds in every model. Then under the proposed semantics, we have the following completeness theorem:

> A ↔ B is a theorem of Angell's system iff it is valid under the truthmaker semantics. (Fine, 2015b; other semantics for which Angell's system is complete are given in Correia, 2004, and Ferguson, 2016.)

This result provides some sort of vindication both of Angell's system and of the proposed account of partial content.

2 Subject-Matter

There is an intuitive notion of subject-matter or of what a statement is about. This notion may have a different focus in different contexts. Thus it may be objectual and concern the objects talked about or it may be predicational and concern what is said about them. Our concern here will be with what one might call 'factual' focus, with what it is in the world that bears upon the statement being true or false.

A standard account of subject-matter, in more or less this sense, was developed by Lewis (1988) and subsequently elaborated by Yablo (2014, ch. 2). A subject-matter, for Lewis, is given by an equivalence relation on worlds where, intuitively, two worlds will stand in the equivalence relation when they do not differ with regard to the subject-matter. Thus if the subject-matter is the current weather in New York, then two worlds will stand in the associated equivalence relation when they do not differ with regard to the current state of the weather in New York.

Lewis has difficulty in defining *the* subject-matter of a statement. Yablo provides a more refined account of subject-matter in terms of truthmakers which removes this difficulty. But he still identifies a subject-matter with a relation on worlds, the subject-matter of a statement now being the similarity relation which holds between two worlds when they share a truthmaker.

There is, I believe, a much more satisfactory and straightforward way of defining subject-matter within the truthmaker framework statement in which no reference is made to worlds or the like (in line with our general unworldly philosophy). Suppose that the verifiers of the statement A are s_1, s_2, …. We may then identify the subject-matter of A with the fusion $\boldsymbol{s} = s_1 \sqcup s_2 \sqcup \ldots$ of its verifiers. After all, it is these states that most directly bear upon the truth of the statement.[3]

It might be thought that this approach to subject-matter is doomed from the start. For consider the statement 'it does or does not rain' and the statement 'it does or does not snow.' Intuitively, their subject-matter is quite different, one concerning the presence or absence of rain and the other the presence or absence of snow. But on our account, the subject-matter of the first statement is the fusion of the presence and absence of rain, let us say, while the subject-matter of the second statement is the fusion of the presence and absence of snow. But these are both impossible states and therefore the same.

But this line of reasoning rests upon adopting the standard coarse-grained conception of impossible states. There is nothing in the truthmaker approach as such which requires us to adopt such a coarse-grained view. Indeed, the more natural view is one in which different

impossible states can be distinguished in terms of the possible states from which they have been obtained (Fine, 2015a); and, in this case, the fine grain of a given subject-matter can be recovered even though its various components have been lumped together into a single impossible state. This then is another case in which impossible states are able to earn their keep.[4]

This account of subject-matter gives rise to a simple and elegant theory of the subject. Since subject-matters are states, we can give an account of the mereology of subject-matters (subject-matter containment, overlap, disjointness, etc.) directly in terms of the mereology of states. We can give a simple account of the subject-matter of complex statements in terms of the subject-matters of their components. Thus where $\sigma(\mathrm{A})$ is the subject-matter of A, we can set:

$$\sigma(\mathrm{A} \wedge \mathrm{B}) = \sigma(\mathrm{A} \vee \mathrm{B}) = \sigma(\mathrm{A}) \sqcup \sigma(\mathrm{B})$$

We can also give a simple account of the *restriction of* a given proposition to some subject-matter $\boldsymbol{s}$. For given two states s and t, let $s \sqcap t$ be their *intersection* (i.e., the fusion of all states that are a common part of s and t). We might then identify the *restriction of* the proposition $P = \{p_1, p_2, \ldots\}$ *to* the subject-matter $\boldsymbol{s}$ to be the proposition $\{p_1 \sqcap \boldsymbol{s}, p_2 \sqcap \boldsymbol{s}, \ldots\}$, obtained by restricting the verifiers of P to the given subject-matter $\boldsymbol{s}$.

Subject-matter will also play a pervasive role in the rest of the theory of truthmakers. Let me give one example from the account of partial content discussed above. When a proposition $P = \{p_1, p_2, \ldots\}$ is closed under fusion, then its subject-matter $\boldsymbol{p} = p_1 \sqcup p_2 \sqcup \ldots$ will itself be a verifier of P. This means that the second clause in the definition of partial content, *viz.* that every verifier of Q should be contained in a verifier of P, can be replaced by the condition that the subject-matter $\boldsymbol{q}$ of Q should be part of the subject-matter $\boldsymbol{p}$ of P.

3 *Counterfactuals*

Ever since the pioneering work of Stalnaker (1968) and Lewis (1973), it has been customary to provide a semantics for counterfactuals statements in terms of possible worlds. The idea, roughly speaking, is to take the counterfactual from A to C to be true just in case the closest world – or all closest worlds, or all sufficiently close worlds – in which A is true is a world – or are worlds – in which C is true. If we introduce a comparative closeness relation $u \preccurlyeq_w v$ (u is as close to w as v), then we may state the third of these options, somewhat more formally, as:

$w \Vdash \mathrm{A} > \mathrm{C}$ if for some world v, $v \Vdash \mathrm{A}$ and, for any world u, $u \Vdash \mathrm{C}$ whenever $u \preceq_w v$ and $u \Vdash \mathrm{A}$.

One familiar difficulty with this account is that it does not enable us to distinguish between the counterfactual 'if Sue were to take the pill then she would live' from the counterfactual 'if Sue were to take the pill or to take the pill and the cyanide then she would live,' even though the first might well be true while the second is false. For the antecedents, 'Sue takes the pill' (p) and 'Sue takes the pill or takes the pill and the cyanide' ($\mathrm{p} \vee (\mathrm{p} \wedge \mathrm{q})$), are truth-functionally equivalent. Indeed, this difficulty is merely the tip of an iceberg. For it can be shown that the possible worlds approach (or any approach that endorses the substitution of truth-functionally equivalent antecedents) is incompatible with our

intuitive judgments about certain scenarios and certain commonly accepted principles of counterfactual reasoning (Fine, 2012b).

In any case, it is worth seeing to what extent these difficulties can be avoided under an alternative approach; and it is here that truthmaker semantics comes into its own. Instead of working with the closeness relation $u \preccurlyeq_w v$ on worlds u, v, and w, we work with a transition relation $t \rightarrow_w u$ on states t and u and world w. Intuitively, $t \rightarrow_w u$ says that u is a possible outcome of imposing t on w. The clause for the truth of a counterfactual at a given world w is then given by:

$w \Vdash A > C$ if for any states t and u for which $t \Vdash A$ and $t \rightarrow_w u$, $u \Vdash C$.[5]

A counterfactual will be true if any exact verifier for the antecedent must transition to an inexact verifier for the consequent. So, for example, the counterfactual 'if the match were struck it would light' will be true because all of the possible outcomes of an exact verifier of 'the match is struck' will be ones which contain an exact verifier for 'the match lights.'

Let us note that this immediately takes care of the Sue example above. Indeed, each of $A > C$ and $B > C$ will be a consequence of $A \vee B > C$ (the so-called rule of Simplification), since the exact verifiers of A and the exact verifies of B are among the exact verifiers of $A \vee B$. So 'if Sue were to take the pill and the cyanide then she would live' will be a consequence of 'if Sue were to take the pill or the pill and cyanide then she would live,' but not of 'If Sue were to take the pill she would live.'

4 *Imperatives*

There seems to be a sense in which one imperative statement may follow from others. Suppose someone says, 'Turn on the light and open the door.' Then it seems that an interlocutor can legitimately say, 'So, turn on the light,' thereby indicating that the one imperative follows from the other. This then raises the question: What is the logic of imperatives? When does one imperative follow from others?

The natural answer to this question is that an imperative inference will be valid just in case the corresponding indicative imperative inference is valid: the imperative Y will follow from the imperatives $X_1, X_2, \ldots, X_n$ just in case the indicative B corresponding to Y follows from the indicatives $A_1, A_2, \ldots, A_n$ corresponding to $X_1, X_2, \ldots, X_n$. Thus in the case above, 'Turn on the light' will follow from 'Turn on the light and shut the door' because 'You turn on the light' follows from 'You turn on the light and shut the door.'

However, this solution does not appear to give the right result in other cases. From 'You turn on the light' follows 'You turn on the light or burn the building down.' But from the imperative 'Turn on the light,' it does not seem as if one can infer 'Turn on the light or burn the building down.' This is Ross's famous paradox.

How then should the semantics and logic of imperatives proceed? Again, it looks as if the framework of truthmaker semantics can provide an answer (Stelzner, 1992; van Rooij, 2000; Aloni and Ciardelli, 2013). Before, we took states to exactly verify or falsify an indicative statement. In the same way, we may take an action (which is a particular kind of state) to be in *exact compliance with* or *exact contravention to* an imperative statement. Thus just as your shutting the door exactly verifies the indicative statement 'You shut the door,' while your shutting the door and turning on the light does not, so your shutting the door is in exact compliance with the imperative statement 'Shut the door,' while your shutting the door and turning on the light is not.

We can then provide compliance and contravention conditions for logically complex imperatives that are the analogue of the verification and falsification conditions for logically complex indicatives. So, for example, we may say:

> the action α is in exact compliance with the conjunctive imperative $X \wedge Y$ iff it is the fusion $\beta \sqcup \gamma$ of an action β that is in exact compliance with X and an action γ that is in exact compliance with Y.

Let the content of an imperative X be the set of actions in exact compliance with the imperative. It may now be suggested that the imperative Y follows from the imperative X just in case the content of Y is part of the content of X. Thus Y will follow from X if (i) any action in compliance with X contains an action in compliance with Y and (ii) any action in compliance with Y will be part of an action in compliance with X. Y must, in this sense, be a necessary means to X.

This account of imperative consequence immediately dissolves Ross's paradox, since the content of $X \vee Y$ will not in general be part of the content of X; and it points to a surprising and intimate connection between the logic of imperatives and the logic of analytic implication.

5 Scalar Implicature

There is a phenomenon that has been discussed in the linguistics literature under the heading 'scalar implicature.' Here is a typical example. Suppose I assert:

(1) John had toast or cereal for breakfast.

Then this is thought to have the 'implicature':

(2) John did not have both toast and cereal for breakfast.

Similarly,

(3) John took one of the candies.

is thought to have the implicature:

(4) John took at most one of the candies.

It has been supposed that, in the first of these cases, there is a sense in which (1) implies (2), since if you believe that John had both toast and cereal then it appears appropriate to register your disagreement with the words:

(5) No (well, in fact), he had both.

But it has also been supposed that the implication is not a regular semantic implication because I can consistently assert:

(6) John had toast or cereal for breakfast and perhaps he even had both.

And likewise in the second case.

How then to account for the implication when it is not a semantic implication? Many philosophers and linguists have appealed to Grice's maxims of cooperative conversation. The rough idea is that if it had been true that John had both toast and cereal for breakfast then I would have said so; and so, from my merely making the weaker claim, you can infer that he did not in fact have both.

However, the Gricean account has a hard time with such examples as:

(1)′ John had toast or cereal or toast and cereal for breakfast; and
(3)′ John took at least one of the candies.

For these do not have the same implicatures – (2) and (4) – as (1) and (3); and yet the same Gricean reasoning would appear to apply.

Let me briefly suggest how these and other difficulties can be avoided under the truthmaker approach, though it will, of course, be necessary to slide over many issues.[6] We suppose that any statement is made against the background of some relevant subject-matter $\boldsymbol{s}$ (again, subject-matter comes into the picture!). Thus if I say that John had toast or cereal for breakfast, we may take the relevant subject-matter to be what he had for breakfast, which we identify with the fusion of various states – his having toast, his having cereal, his having bacon and eggs, and so on.

Any subject-matter $\boldsymbol{s}$ will have an actual part $\boldsymbol{s}_{@}$, which is the fusion of all those parts of $\boldsymbol{s}$ that are actual. Where $\boldsymbol{s}$ is the relevant subject-matter of a statement A, we may now say that A is *exactly true* if $\boldsymbol{s}_{@}$ is an exact verifier of A. We might call $\boldsymbol{s}_{@}$, when $\boldsymbol{s}$ is the relevant subject-matter, the *relevant situation*. Then for A to be exactly true is for it to be exactly verified by the relevant situation.

We make the following key hypothesis:

> The Origin of Scalar Implicature (OSI): Scalar implicature arises from the presupposition that a statement is not merely true but exactly true.

What we say should, in this sense, fit the facts. Let us now see how this hypothesis can take care of the difficulties mentioned above. Consider (1). If John had both cereal and toast for breakfast, then the relevant situation would involve his having both for breakfast and so would not be an exact verifier of (1). Thus the exact truth of (1) requires that John not have both cereal and toast for breakfast. Similarly for (3). If John took more than one candy, then the relevant situation would involve his taking several candies and so would not be an exact verifier of (3) (which should be the verifier of a single instance of (3)). Thus the exact truth of (3) requires that John take a single candy.

Contrast this now with (1)′ and (3)′. In this case, John having both cereal and toast would be an exact verifier of (1)′ and so the exact truth of (1)′ does not require the truth of (2). Similarly for (3)′. For in this case, we may suppose that John taking several candies is an exact verifier for (3)′ (in line with the inclusive semantics for the existential quantifier in §I7) and so again the implicature will be lost.

In this chapter, I have provided the merest sketch of the truthmaker approach. The abstract theory may be developed in many other directions and many other examples of its application might be given: to ground (Correia, 2010; Fine, 2012d; 2015c), the determinate/determinable distinction (Fine, 2011), and the status of impossible states (Fine, 2015a)

within metaphysics; to modal and deontic logic, intuitionistic logic (Fine, 2014a), and the theory of logical remainder and common content (Fine, 2015c; Yablo, 2014, ch. 8) within philosophical logic; to quantification, anaphora, free choice disjunction, intensional descriptions (Moltmann, 2017), cases-constructions (Moltmann, 2015), adverbial modification (van Fraassen, 1973; Hwang and Schubert, 1993), presupposition (Yablo, 2014, ch. 10), the logic and semantics of questions (Ciardelli, Groenendijk, and Roelefson, 2013), and vagueness (van Rooij, 2013) within the philosophy of language and linguistics; to belief revision and closure principles for knowledge and belief (Yablo, 2014, ch. 7) within epistemology; to verisimilitude (Fine, 2015e; Gemes, 2007; Yablo, 2014, §6.7), confirmation (Gemes, 1994; 1997; Yablo, 2014, §§6.1–6.5), and causal modeling within the philosophy of science; and to the psychology of reasoning (Koralus and Mascarenhas, 2013) and the frame problem within cognitive science. But I hope I have said enough to give the reader a taste of what the theory is like, of what it is capable of doing, and of how, in many respects, it is far superior to the more usual approach in terms of possible worlds.

Notes

1 One might also provide a truthmaker semantics for *intuitionistic* sentential logic (Fine, 2014a) and, in this case, it is not necessary to state separate clauses for the verification and falsification of an arbitrary statement.

2 Allowing impossible totality states of this sort greatly simplifies the semantics and is another example of the benefits to be gained by admitting impossible states.

3 Fine (2015d). Strictly speaking, s is the *positive* subject-matter of A. The *negative* subject-matter may be taken to be $t = t_1 \sqcup t_2 \sqcup \ldots$, where t_1, t_2, … are the falsifiers of A; and the *overall* subject-matter may be taken to be (s, t) or $s \sqcup t$.

4 A further case is the truthmaker semantics for intuitionistic logic in Fine (2014a), which also makes use of a rich ontology of impossible states.

5 Fine (2012c). We may also give the following more complicated clause for the exact verification of a counterfactual:

$s \Vdash A > C$ iff (i) for each t for which $t \Vdash A$ there is an s_t and a u for which $t \rightarrow s_t\, u$,
(ii) for any state t for which $t \Vdash A$ and any state u for which $t \rightarrow s_t\, u$, $u \Vdash > C$,
(iii) s is the fusion of $\{s_t\colon t \Vdash A\}$;

and give a related clause for the exact falsification of a counterfactual.

6 I first publicly presented a solution along these lines in the Jack Smart lecture of 2010; and a similar solution has been independently proposed by Robert van Rooij in his (2013) and in some recent unpublished work.

References

Aloni, M., M. Franke, and F. Roelofsen, eds. 2013. *The Dynamic, Inquisitive, and Visionary Life of φ, ?φ, and ◊φ: Festschrift for Jeroen Groenendijk, Martin Stokhof, and Frank Veltman*. Netherlands: Onbekend.

Aloni, M., and I. Ciardelli. 2013. "A logical account of free choice imperatives." In Aloni, Franke, and Roelofsen, 2013, pp. 1–17.

Angell, R. B. 1977. "Three systems of first degree entailment." *Journal of Symbolic Logic*, 42(1): 147.

Angell, R. B. 1989. "Deducibility, entailment and analytic containment." In *Directions in Relevant Logic*, edited by J. Norman and R. Sylvan, pp. 119–144. Dordrecht, Netherlands: Kluwer Academic Publishers.

Angell, R. B. 2002. *A-Logic*. Lanham, MD: University Press of America.

Armstrong, D. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Armstrong, D. 2004. *Truth and Truthmakers*. Cambridge: Cambridge University Press.

Barwise, J., and J. Perry. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.

Black, M., and P. Geach, eds. 1980. *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edn. Oxford: Blackwell.

Ciardelli, I., J. Groenendijk, and F. Roelefson. 2013. "Inquisitive semantics: a new notion of meaning." *Compass*, 7(9): 459–476.

Correia, F. 2004. "Semantics for analytic containment." *Studia Logica*, 77(1): 87–104.

Correia, F. 2010. "Grounding and truth-functions." *Logique et Analyse*, 53: 251–279.

Correia, F., and B. Schnieder, eds. 2012. *Metaphysical Grounding*. Cambridge: Cambridge University Press.

Davidson, D. 1967. "Truth and meaning." *Synthese*, 17: 304–323.

Ferguson, T. 2016. "Faulty Belnap computers and subsystems of FDE." *Journal of Logic and Computation*, 26(5): 1617–1636.

Fine, K. 1975. "Review of David Lewis' *Counterfactuals*." *Mind*, 84: 451–458. Reprinted in Fine, 2005, ch.10.

Fine, K. 1985. *Reasoning with Arbitrary Objects*. Blackwell: Oxford.

Fine, K. 2005. *Modality and Tense*. Oxford: Oxford University Press.

Fine, K. 2011. "An abstract characterization of the determinate/determinable distinction." *Philosophical Perspectives*, 25: 161–187.

Fine, K. 2012a. "Guide to ground." In *Metaphysical Grounding*, edited by F. Correia and B. Schneider, pp. 37–80. Cambridge: Cambridge University Press.

Fine, K. 2012b. "A difficulty for the possible worlds analysis of counterfactuals." *Synthese*, 289(1): 29–57.

Fine, K. 2012c. "Counterfactuals without possible worlds." *Journal of Philosophy*, 109(3): 221–246.

Fine, K. 2012d. "The pure logic of ground." *Review of Symbolic Logic*, 25(1): 1–25.

Fine, K. 2013. "A note on partial content." *Analysis*, 73(3): 413–419.

Fine, K. 2014a. "Truthmaker semantics for intuitionistic logic." *Journal of Philosophical Logic*, 43(2): 549–577. Reprinted in *Philosophers' Annual*, vol. 33. http://www.philosophersannual.org/ (accessed August 24, 2016).

Fine, K. 2014b. "Permission and possible worlds." *Dialectica*, 68(3): 317–336.

Fine, K. 2015a. "Constructing the impossible." In a collection of papers for Dorothy Edgington, forthcoming.

Fine, K. 2015b. "Truthconditional content I." Forthcoming – to appear in *Journal of Philosophical Logic*.

Fine, K. 2015c. "Truthconditional content II." Forthcoming – to appear in *Journal of Philosophical Logic*.

Fine, K. 2015d. "Partial Truth." Forthcoming.

Fine, K. 2015e. "Verisimilitude." Forthcoming.

Fine, K. 2016. "Angellic content." *Journal of Philosophical Logic*, 45(2): 199–226.

Frege, G. 1892, "Über Sinn und Bedeutung." In *Zeitschrift für Philosophie und Philosophische Kritik*, 100: 25–50. Translated as "On Sense and Reference" in Black and Geach, 1980, pp. 56–78.

Gemes, K. 1994. "A new theory of content I: basic content." *Journal of Philosophical Logic*, 23(6): 595–620.

Gemes, K. 1997. "A new theory of content II: model theory and some alternatives." *Journal of Philosophical Logic*, 26(4): 449–476.

Gemes, K. 2007. "Verisimilitude and content." *Synthese*, 154(2): 293–306.

Groenendijk, J., and F. Roelofsen. 2010. "Radical inquisitive semantics." http://www.illc.uva.nl/inquisitive-semantics (accessed August 24, 2016).

Horwich, P. 2008. "Being and truth." *Midwest Studies in Philosophy*, 32: 258–273.

Humberstone, L. 1981. "From worlds to possibilities." *Journal of Philosophical Logic*, 10(1): 313–339.

Hwang, C. H., and L. K. Schubert. 1993. "Episodic logic: a situational logic for natural language processing." In *Situation Theory and its Applications 3 (STA-3)*, edited by P. Aczel, D. Israel, Y. Katagiri, and S. Peters, pp. 307–452. Stanford, CA: Center for the Study of Language and Information.

Koralus P., and S. Mascarenhas. 2013. "The erotetic theory of reasoning: bridges between formal semantics and the psychology of propositional deductive inference." *Philosophical Perspectives*, 27: 312–365.

Kratzer, A. 2014. "Situations in natural language semantics." *Stanford Encylopedia of Philosophy*. http://plato.stanford.edu/entries/situations-semantics/ (accessed August 24, 2016).

Lewis, D. K. 1973. *Counterfactuals*. Oxford: Blackwell.

Lewis, D. K. 1988. "Statements partly about observation." In *Papers in Philosophical Logic*, pp. 125–155. Cambridge: Cambridge University Press.

Moltmann, F. 2007. "Events, tropes and truthmaking." *Philosophical Studies*, 134(3): 363–403.

Moltmann, F. 2015. "A truthmaker semantics for cases." In *Philosophical Research Online*, http://semanticsarchive.net/Archive/DZhMGVjZ/cases.pdf (accessed October 6, 2016).

Moltmann, F. 2017. "Variable objects and truthmaking." In *Metaphysics, Meaning and Modality: Themes from Kit Fine*, edited by M. Mircea. Oxford: Oxford University Press.

Montague, R. 1970. "English as a formal language." In *Linguaggi nella Società et nella Tecnica*, edited by B. Visentini *et al.,* pp. 188–221. Milan: Edizioni di Communità.

Rumfitt, I. 2012. "A neglected path to intuitionism." *Topoi*, 31(1): 101–109.

Russell, B. 1918/1919. "The philosophy of logical atomism." *The Monist*, 28: 495–527; 29: 190–222, 345–380.

Schnieder, B. 2006. "Troubles with truth-making: necessitation and projection." *Erkenntnis*, 64(1): 61–74.

Stalnaker, R. 1968. "A theory of conditionals." In *Studies in Logical Theory, American Philosophical Quarterly Monograph Series*, no. 2, edited by N. Rescher, pp. 98–112. Oxford: Blackwell.

Stelzner, W. 1992. "Relevant deontic logic." *Journal of Philosophical Logic*, 21(2): 193–216.

van Fraassen, B. 1969. "Facts and tautological entailments." *Journal of Philosophy*, 66(15): 477–487.

van Fraassen, B. 1973. "Extension, intension and comprehension." In *Logic and Ontology*, edited by M. Munitz, pp. 101–131. New York: New York University Press.

van Rooij, R. 2000. "Permission to change." *Journal of Semantics*, 17(2): 119–145.

van Rooij, R. 2013. "Vagueness: insights from Martin, Jeroen and Frank." In Aloni, Franke, and Roelofsen, 2013, pp. 216–228.

Yablo, S. 2014. *Aboutness*. Princeton, NJ: Princeton University Press.

23

Analyticity

PAUL ARTIN BOGHOSSIAN

I

This is what many philosophers believe today about the analytic/synthetic distinction: In his classic early writings on analyticity – in particular, in "Truth by convention" (1976a), "Two dogmas of empiricism" (1953), and "Carnap and logical truth" (1976b) – Quine showed that there can be no distinction between sentences that are true purely by virtue of their meaning, and those that aren't. In so doing, Quine devastated the philosophical programs that depend upon a notion of analyticity – specifically, the linguistic theory of necessary truth, and the analytic theory of *a priori* knowledge.

Quine himself, so the story continues, went on to espouse far more radical views about meaning, including such theses as meaning-indeterminacy and meaning-skepticism. However, it is not necessary, and certainly not appealing, to follow him on this trajectory. As realists about meaning, we may treat Quine's self-contained discussion in the early papers as the basis for a profound *insight* into the nature of meaning-facts, rather than for any sort of rejection of them. We may discard the notions of the analytic and the *a priori* without thereby buying in on any sort of unpalatable skepticism about meaning.

Now, I don't know precisely how many philosophers believe all of the above, but I think it would be fair to say that it is the prevailing view. Philosophers with radically differing commitments – including radically differing commitments about the nature of meaning itself – subscribe to it: whatever precisely the correct construal of meaning, so they seem to think, Quine has shown that it will not sustain a distinction between the analytic and the synthetic. Here, merely for purposes of illustration, are two representative endorsements of the view, both of them also containing helpful references to its popularity. The first is by Bill Lycan.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

> It has been nearly forty years since the publication of "Two Dogmas of Empiricism." Despite some vigorous rebuttals during that period, Quine's rejection of analyticity still prevails – in that philosophers *en masse* have either joined Quine in repudiating the "analytic/synthetic" distinction or remained (however mutinously) silent and made no claims of analyticity.
>
> This comprehensive capitulation is somewhat surprising, in light of the radical nature of Quine's views on linguistic meaning generally. In particular, I doubt that many philosophers accept his doctrine of the indeterminacy of translation.

Lycan goes on to promise that, in his paper, he is going to

> make a Quinean case against analyticity, without relying on the indeterminacy doctrine. For I join the majority in denying both analyticity and indeterminacy. (Lycan, 1991, p. 111)

Next, here are two other committed realists about meaning, Jerry Fodor and Ernie Lepore, talking about a thesis that, they say,

> practically everybody thinks that there are good reasons to endorse;… [namely]… that there aren't any expressions that are true or false solely in virtue of what they mean. (Fodor and Lepore, 1991a, p. 332)

Fodor and Lepore go on to claim that this result clearly undermines the idea of a belief or inference that is warranted *a priori*.

Now, my disagreement with the prevailing view is not total. There is *a* notion of 'truth by virtue of meaning' – what I shall call the metaphysical notion – that *is* undermined by a set of indeterminacy-independent considerations. Since this notion is presupposed by the linguistic theory of necessity, that project fails and must be abandoned.

However, I disagree with the prevailing view's assumption that those very same considerations also undermine the analytic explanation of the *a priori*. For I believe that an entirely distinct notion of analyticity underlies that explanation, a notion that is epistemic in character. And in contrast with the metaphysical notion, the epistemic notion can be defended, I believe, provided that even a minimal realism about meaning is true. I'm inclined to hold, therefore, that there can be no effective Quinean critique of the *a priori* that does not ultimately depend on Quine's radical thesis of the indeterminacy of meaning, a thesis that, as I've stressed, many philosophers continue to reject.

All of this is what I propose to argue in this chapter. I should emphasize right at the outset, however, that I am not a historian, and my interest here is not historical. Think of me, rather, as asking, on behalf of all those who continue to reject Quine's later skepticism about meaning: Can something like the analytic explanation of the *a priori* be salvaged from the wreckage of the linguistic theory of necessity?

Belief, Apriority, and Indeterminacy

We need to begin with some understanding, however brief and informal, of what it is to believe something, and of what it is for a belief to count as *a priori* knowledge.

In my view, the most plausible account of the matter is that believing is a relation to a proposition in the technical sense: a mind-independent, language-independent abstract

object that has its truth-conditions essentially. Against this background, a belief is true just in case its proposition is true.

However, I don't want to presuppose such a picture of belief in the present context. Not that there would be anything particularly wrong or question-begging about doing so; as Quine himself has made clear, his rejection of propositions is supposed to rest on his critique of analyticity, not the other way around.[1] Nevertheless, in the interests of keeping potential distractions to a minimum, I will work with a picture of belief that is far more hospitable to Quine's basic outlook.

According to this more 'linguistic' picture, the objects of belief are not propositions, but rather interpreted sentences: for a person T to believe that p is for T to hold true a sentence S which means that p in T's idiolect.[2]

Against this rough and ready background, we may say that for T to know that p is for T to justifiably hold S true, with a strength sufficient for knowledge, and for S to be true. And to say that T knows p *a priori* is to say that T's warrant for holding S true is independent of outer, sensory experience.[3] The interesting question in the analysis of the concept of apriority concerns this notion of warrant: What is it for a belief to be justified, independently of outer sensory experience?

On a minimalist reading, to say that the warrant for a given belief is *a priori* is just to say that it is justified, with a strength sufficient for knowledge, without appeal to empirical evidence.[4] On a stronger reading, it is to say that, and to say in addition that the justification in question is not defeasible by any future empirical evidence.[5] Which of these two notions is at issue in the present debate?

My own view is that the minimal notion forms the core of the idea of apriority. However, in this chapter I will aim to provide the materials with which to substantiate the claim that, under the appropriate circumstances, the notion of analyticity can help explain how we might have *a priori* knowledge even in the strong sense. A defense of the strong notion is particularly relevant in the present context, for Quine seems to have been particularly skeptical of the idea of empirical indefeasibility.

Before proceeding, we should also touch briefly on the notion of meaning-indeterminacy. In chapter 2 of *Word and Object* Quine (1960) argued that, for any language, it is possible to find two incompatible translation manuals that nevertheless perfectly conform to the totality of the evidence that constrains translation. This is the famous doctrine of the indeterminacy of translation. Since Quine was, furthermore, prepared to assume that there could not be facts about meaning that are not captured in the constraints on best translation, he concluded that meaning-facts themselves are indeterminate – that there is, strictly speaking, no determinate fact of the matter as to what a given expression in a language means. This is the doctrine that I have called the thesis of the indeterminacy of meaning.

An *acceptance* of meaning-indeterminacy can lead to a variety of *other* views about meaning. For instance, it might lead to an outright eliminativism about meaning. Or it might be taken as a reason to base the theory of meaning on the notion of likeness of meaning, rather than on that of sameness of meaning (see Harman, 1973). In this chapter I am not concerned with the question of what moral should be drawn from the indeterminacy thesis, on the assumption that it is true; nor am I concerned with whether the indeterminacy thesis is true. I am only concerned to show that a skepticism about epistemic analyticity cannot stop short of the indeterminacy thesis, a thesis that, as I have stressed, most philosophers agree in rejecting (see Chapter 26, INDETERMINACY OF TRANSLATION).

Analyticity: Metaphysical or Epistemological?

Traditionally, three classes of statements have been thought to be the objects of *a priori* knowledge: logical statements, exemplified by such truths as:

Either Brutus killed Caesar or he did not;

mathematical statements, such as:

$7 + 5 = 12$;

and conceptual truths, for instance:

All bachelors are unmarried.

The problem has always been to explain how any statement could be known *a priori*. After all, if a statement is known *a priori*, then it must be true. And if it is true, then it must be factual, capable of being true or false. What could possibly entitle us to hold a factual sentence true on *a priori* grounds?

The history of philosophy has known a number of answers to this question, among which the following has had considerable influence: We are equipped with a special evidence-gathering faculty of *intuition*, distinct from the standard five senses, which allows us to arrive at justified beliefs about the necessary properties of the world. By exercising this faculty, we are able to know *a priori* such truths as those of mathematics and logic.

The central impetus behind the *analytic* explanation of the *a priori* is the desire to explain the possibility of *a priori* knowledge without having to postulate such a special faculty, one that has never been described in satisfactory terms. The question is: How could a factual statement S be known *a priori* by T, without the help of a special evidence-gathering faculty?

Here, it would seem, is one way: *If mere grasp of* S*'s meaning by* T *sufficed for* T*'s being justified in holding* S *true.* If S were analytic in this sense, then, clearly, its apriority would be explainable without appeal to a special faculty of intuition: mere grasp of its meaning by T would suffice for explaining T's justification for holding S true. On this understanding, then, 'analyticity' is an overtly *epistemological* notion: a statement is 'true by virtue of its meaning' provided that grasp of its meaning alone suffices for justified belief in its truth.

Another, far more metaphysical, reading of the phrase 'true by virtue of meaning' is also available, however, according to which a statement is analytic provided that, in some appropriate sense, it *owes its truth-value completely to its meaning*, and not at all to 'the facts.'

Which of these two possible notions has been at stake in the dispute over analyticity? There has been a serious unclarity on the matter. Quine himself tends to label the doctrine of analyticity an epistemological one, as, for example, in the following passage from "Carnap and logical truth":

> the linguistic doctrine of logical truth, which is an epistemological doctrine, goes on to say that logical truths are true purely by virtue of the intended meanings, or intended usage, of the logical words. (Quine, 1976b, p. 103)

However, his most biting criticisms seem often to be directed at what I have called the metaphysical notion. Consider, for example, the object of disapproval in the following famous passage, a passage that concludes the discussion of analyticity in "Two dogmas":

> It is obvious that truth in general depends on both language and extralinguistic fact. The statement 'Brutus killed Caesar' would be false if the world had been different in certain ways, but it would also be false if the word 'killed' happened rather to have the sense of 'begat.' Thus one is tempted to suppose in general that the truth of a statement is somehow analyzable into a linguistic component and a factual component. Given this supposition it next seems reasonable that in some statements the factual component should be null; and these are the analytic statements. But for all its a priori reasonableness, a boundary between analytic and synthetic statements simply has not been drawn. That there is such a distinction to be drawn at all is an unempirical dogma of empiricists, a metaphysical article of faith. (Quine, 1953, pp. 36–37)

Now, I think that there is no doubt that many of the proponents of the analytic theory of the *a priori*, among them especially its positivist proponents, intended the notion of analyticity to be understood in this metaphysical sense; very shortly I shall look at why.

Before doing that, however, I want to register my wholehearted agreement with Quine, that the metaphysical notion is of dubious explanatory value, and possibly also of dubious coherence. I believe that Quine's discrediting of this idea constitutes one of his most enduring contributions to philosophy. Fortunately for the analytic theory of the *a priori*, it can be shown that it need have nothing to do with the discredited idea.

The Metaphysical Concept

What could it possibly mean to say that the truth of a statement is fixed exclusively by its meaning and not by the facts? Isn't it in general true – indeed, isn't it in general a truism – that for any statement S,

> S is true iff for some p, S means that p and p?

How could the *mere* fact that S means that p make it the case that S is true? Doesn't it also have to be the case that p? As Harman has usefully put it (he is discussing the sentence 'Copper is copper'):

> what is to prevent us from saying that the truth expressed by "Copper is copper" depends in part on a general feature of the way the world is, namely that everything is self-identical. (Harman, 1968, p. 128)[6]

The proponent of the metaphysical notion does have a comeback, one that has perhaps not been sufficiently addressed. If he is wise, he won't want to deny the meaning-truth truism. What he will want to say instead is that, in some appropriate sense, our meaning p by S *makes it the case that p*.

But this line is itself fraught with difficulty. For how can we make sense of the idea that something is made true by our meaning something by a sentence?

Consider the sentence 'Either p or not p.' It is easy, of course, to understand how the fact that we mean what we do by the ingredient terms fixes what is expressed by the sentence as

a whole; and it is easy to understand, in consequence, how the fact that we mean what we do by the sentence determines whether the sentence expresses something true or false. But as Quine points out, that is just the normal dependence of truth on meaning. What is far more mysterious is the claim that the *truth of what the sentence expresses* depends on the fact that it is expressed by that sentence, so that we can say that what is expressed wouldn't have been true at all, had it not been for the fact that it is expressed by that sentence. There are at least two insurmountable problems in making sense of this idea.

First, any such account would make the truth of what is expressed *contingent*, whereas most of the statements at stake in the present discussion are clearly necessary. Second, such an account would make the truth of the claim expressed contingent *on* an act of meaning, and that is very peculiar. Putting aside the question whether it is so much as intelligible, what plausibility could it conceivably have? Are we to suppose that, prior to our stipulating a meaning for the sentence

Either snow is white or it isn't

it wasn't the case that either snow was white or it wasn't? Isn't it overwhelmingly obvious that this claim was true *before* such an act of meaning, and that it would have been true even if no one had thought about it, or chosen it to be expressed by one of our sentences?

Why, if this idea is as problematic as I, following Quine, have claimed it to be, did it figure so prominently in positivist thinking about analyticity?

Part of the answer derives from the fact that the positivists didn't merely want a theory of *a priori* knowledge; they also wanted a reductive theory of necessity. The motivation was not purely epistemological, but metaphysical as well. Guided by the fear that objective, language-independent, necessary connections would be metaphysically odd, they attempted to show that all necessities could be understood to consist in linguistic necessities, in the shadows cast by conventional decisions concerning the meanings of words. Linguistic meaning, by itself, was supposed to generate necessary truth; *a fortiori*, linguistic meaning, by itself, was supposed to generate truth. Hence the play with the metaphysical concept of analyticity.

But this is, I believe, a futile project. In general, I have no idea what would constitute a better answer to the question: 'What is responsible for generating the truth of a given class of statements?' than something bland like 'the world' or 'the facts'; and, for reasons that I have just been outlining, I cannot see how a good answer might be framed in terms of meaning in particular.

So I have no sympathy with the linguistic theory of necessity or with its attendant Conventionalism. Unfortunately, the impression appears to be widespread that there is no way to disentangle that view from the analytic theory of the *a priori*; or, at a minimum, that there is no way to embrace the epistemic concept of analyticity without also embracing its metaphysical counterpart. I don't know whether Harman believes something of the sort; he certainly gives the impression of doing so in his frequent suggestions that anyone deploying the notion of analyticity would have to be deploying both of its available readings simultaneously:

> It turned out that someone could be taught to make the analytic-synthetic distinction only by being taught a rather substantial theory, a theory including such principles as that meaning can make something true and that knowledge of meaning can give knowledge of truth. (Harman, 1994a, p. 47; see also Harman, 1968)

One of the main points of the present chapter is that these two notions of analyticity are distinct, and that the analytic theory of the *a priori* needs only the epistemological notion and has no use whatsoever for the metaphysical one. We can have an analytic theory of the *a priori* without in any way subscribing to a Conventionalism about anything. It is with the extended defense of this claim that much of the present chapter is concerned.

The Epistemological Concept

Turning, then, to the epistemological notion of analyticity, we immediately confront a serious puzzle: How could any sentence be analytic in this sense? How could mere grasp of a sentence's meaning justify someone in holding it true?

Clearly, the answer to this question has to be *semantical*: something about the sentence's meaning, or about the way that meaning is fixed, must explain how its truth is knowable in this special way. What could this explanation be?

In the history of the subject, two different sorts of explanation have been especially important. Although these, too, have often been conflated, it is crucial to distinguish between them.

One idea was first formulated in full generality by Gottlob Frege. According to Frege, a statement's analyticity (in my epistemological sense) is to be explained by the fact that it is *transformable into a logical truth by the substitution of synonyms for synonyms*. When a statement satisfies this semantical condition, I shall say that it is 'Frege-analytic.'[7]

Now, it should be obvious that Frege-analyticity is at best an *incomplete* explanation of a statement's epistemic analyticity and, hence, of its apriority. For suppose that a given sentence S is Frege-analytic. How might this fact explain its analyticity? Clearly, two further assumptions are needed. First, that facts about synonymy are knowable *a priori*; and second, that so are the truths of logic. Under the terms of these further assumptions, a satisfying explanation goes through. Given its Frege-analyticity, S is transformable into a logical truth by the substitution of synonyms for synonyms. Facts about synonymy are *a priori*, so it's *a priori* that S is so transformable. Furthermore, the sentence into which it is transformable is one whose truth is itself knowable *a priori*. Hence, S's truth is knowable *a priori*.

Frege tended not to worry about these further assumptions for two reasons. First, he thought it obviously constitutive of the idea of meaning that meaning is transparent – that any competent user of two words would have to be able to know *a priori* whether or not they meant the same. Second, he also thought it obvious that there could be no substantive epistemology for logic – *a fortiori*, not one that could explain its apriority. As a consequence, he was happy to take logic's apriority for granted. For both of these reasons, he didn't worry about the fact that the concept of Frege-analyticity simply leaned on these further assumptions without explaining them.

I think the jury is still out on whether Frege was right to take these further assumptions for granted. There is certainly a very strong case to be made for the transparency of meaning.[8] And there are well-known difficulties providing a substantive epistemology for something as basic as logic, difficulties we shall have occasion to further review below. Nevertheless, because we cannot simply assume that Frege was right, we have to ask how a complete theory of the *a priori* would go about filling in the gaps left by the concept of Frege-analyticity.

I shall have very little to say about the first gap. The question whether facts about the sameness and difference of meaning are *a priori* cannot be discussed independently of the question of what meaning is, and that is not an issue that I want to prejudge in the present context. On some views of meaning – for example, on certain conceptual-role views – the apriority of synonymy is simply a by-product of the very nature of meaning-facts, so that no substantive epistemology for synonymy is necessary or, indeed, possible. On other views – for example, on most externalist views of meaning – synonymy is not *a priori*, so there is no question of a sentence's Frege-analyticity fully explaining its epistemic analyticity.

Since this issue about the apriority of synonymy turns on questions that are currently unresolved, I propose to leave it for now. As we shall see, none of the analyticity-skeptical considerations we shall consider exploit it in any way. (Quine never argues that the trouble with Frege-analyticity is that synonymies are *a posteriori*.)

Putting aside, then, skepticism about the apriority of synonymy, and, for the moment anyway, skepticism about the very existence of Frege-analytic sentences, let us ask quite generally: What class of *a priori* statement would an account based on the notion of Frege-analyticity *fail* to explain?

Two classes come to mind. On the one hand, *a priori* statements that are not transformable into logical truths by the substitution of synonyms for synonyms; and, on the other, *a priori* statements that are trivially so transformable.

Taking the first class first, there does appear to be a significant number of *a priori* statements that are not Frege-analytic. For example:

Whatever is red all over is not blue.
Whatever is colored is extended.
If x is warmer than y, then y is not warmer than x.

These statements appear not to be transformable into logical truths by the appropriate substitutions: the ingredient descriptive terms seem not to be decomposable in the appropriate way.

The second class of recalcitrant statements consists precisely of the truths of logic. The truths of logic satisfy, of course, the conditions on Frege-analyticity: but they satisfy them trivially. And it seems obvious that we can't hope to explain our entitlement to belief in the truths of logic by appealing to their analyticity in this sense: knowledge of Frege-analyticity presupposes knowledge of logical truth, and so can't explain it.

How, then, is the epistemic analyticity of these recalcitrant truths to be explained? As we shall see below, the Carnap/Wittgenstein solution turned on the suggestion that they are to be viewed as *implicit definitions* of their ingredient terms. When a statement satisfies this semantical condition, I shall sometimes say that it is 'Carnap-analytic.' However, before proceeding to a discussion of Carnap-analyticity I want to re-examine Quine's famous rejection of the much weaker concept of Frege-analyticity.

II

"Two Dogmas" and the Rejection of Frege-Analyticity

For all its apparent limitations, the concept of Frege-analyticity is not without interest. Even though Quine made it fashionable to claim otherwise, "All bachelors are male" *does* seem to be transformable into a logical truth by the substitution of synonyms for synonyms, and

that fact *does* seem to have something important to do with that statement's apriority. If, then, appearances are not misleading here, and a significant range of *a priori* statements are Frege-analytic, then the problem of their apriority is *reduced* to that of the apriority of logic and synonymy and, in this way, a significant economy in explanatory burden is achieved.

It was, therefore, an important threat to the analytic theory of the *a priori* to find Quine arguing, in one of the most celebrated articles of the twentieth century, that the apriority of no sentence could be explained by appeal to its Frege-analyticity, because no sentence of a natural language could *be* Frege-analytic.

It has not been sufficiently appreciated, it seems to me, that "Two dogmas" is *exclusively* concerned with this weaker notion of Frege-analyticity, and not at all with the more demanding project of explaining the apriority of logic. But this is made very clear by Quine:

> Statements which are analytic by general philosophical acclaim are not, indeed, far to seek. They fall into two classes. Those of the first class, which may be called *logically true*, are typified by:
>
> (1) No unmarried man is married.
>
> The relevant feature of this example is that it is not merely true as it stands, but remains true under any and all reinterpretations of 'man' and 'married.' If we suppose a prior inventory of *logical* particles … then in general a logical truth is a statement that remains true under all reinterpretations of its components other than the logical particles.
>
> But there is also a second class of analytic statements, typified by:
>
> (2) No bachelor is married.
>
> The characteristic of such a statement is that it can be turned into a logical truth by putting synonyms for synonyms. (1953, pp. 22–23)

Quine goes on to say very clearly:

> Our problem … is analyticity; and here the major difficulty lies not in the first class of analytic statements, the logical truths, but rather in the second class, which depends on the notion of synonymy. (1953, p. 24)

Most of the rest of "Two dogmas" is devoted to arguing that no good sense can be made of such analyticities of the 'second class.'

None of this would make any sense unless Quine were intending in "Two dogmas" to be restricting himself solely to the notion of Frege-analyticity. Of course, it is the point of two other important papers of his – "Truth by convention" and "Carnap and logical truth" – to argue that there is no non-trivial sense in which *logic* is analytic. We will turn to that issue in due course. Relative to the Fregean notion, however, the logical truths are trivially analytic; and so, given his apparent desire to restrict his attention to that notion in "Two dogmas," he simply concedes their 'analyticity' in the only sense he takes to be under discussion. What he wishes to resist in "Two dogmas," he insists, is merely the claim that there are any *non-trivial instances of Frege-analyticity*.[9]

Skeptical Theses about Analyticity

What form does Quine's resistance take? Let's agree, right away, that the result being advertised isn't anything modest, of the form: There are fewer analyticities than we had previously thought. Or, there are some analytic truths, but they are not important for the

purposes of science. Or anything else of a similar ilk. Rather, as a very large number of Quine's remarks make clear, the sought-after result is something ambitious, to the effect that the notion of Frege-analyticity is, somehow or other, not cogent. The many admirers of "Two dogmas" have been divided on whether to read this as the claim that the notion of Frege-analyticity does not have a well-defined, determinate content, or whether to read it merely as claiming that, although it has an intelligible content, it is necessarily uninstantiated.

I'll call the first claim a *Non-factualism* about analyticity:

> (NF) No coherent, determinate property is expressed by the predicate 'is analytic' (or, since these are correlative terms, the predicate 'is synthetic'); consequently, no coherent proposition is expressed by sentences of the form 'S is analytic' and 'S is synthetic.'

And I'll call the second an *Error Thesis* about analyticity:

> (ET) There is a coherent, determinate property expressed by 'is analytic,' but it is necessarily uninstantiated; consequently, all sentences of the form 'S is analytic' are necessarily false.[10]

Unfortunately, "Two dogmas" doesn't seem to have a clear view about exactly which of these claims it should be read as arguing for.

In favor of the suggestion that Quine's goal is something with the form of a non-factualism about Frege-analyticity there is, first, the fact that the idiom favored by Quine – that there is no distinction between the analytic and the synthetic – sits much better with a non-factualist thesis than it does with an error thesis. The latter claim would be far more happily expressed by saying, "All sentences are necessarily synthetic."

Further, and more importantly, there is the actual character of Quine's *arguments*. As any reader of "Two dogmas" knows, much of that article is given over to arguing that we don't really understand what 'is analytic' means, that previous explications either fail to specify its meaning in sufficiently non-circular – hence sufficiently illuminating – terms, or fail to specify it at all.

For example, against the suggestion that 'analyticity' might be understood via a specification of the 'semantical rules' for a language, Quine remarks:

> Let us suppose … an artificial language L_o whose semantical rules have the form explicitly of a specification, by recursion or otherwise, of all the analytic statements of L_o. The rules tell us that such-and-such statements, and only those, are the analytic statements of L_o. Now here the difficulty is simply that the rules contain the word 'analytic' which we do not understand! We understand what expressions the rules attribute analyticity to, but we do not understand what the rules attribute to these expressions. (Quine, 1953, p. 33)

There are, then, weighty textual reasons for taking Quine to be arguing for something with the form of an NF. Other considerations, however, pull in the opposite direction. The most striking of these occurs in the following passage concerning stipulative definitions, that is, the explicitly conventional introduction of novel notation for the purposes of abbreviation. The passage is framed by a concession on Quine's part that Frege-analyticity would be intelligible, provided the notion of synonymy were. In the case of stipulative definitions, writes Quine,

> the definiendum becomes synonymous with the definiens simply because it has been created expressly for the purpose of being synonymous with the definiens. Here we have a really transparent case of synonymy created by definition; would that all species of synonymy were as intelligible. (Quine, 1953, p. 26)

This admission, however, in the context of Quine's concession, would appear to be utterly inconsistent with NF. For an NF about Frege-analyticity is committed to the claim that there is no coherent, determinate property of synonymy: no conceivable mechanism could generate an instance of synonymy, for there is no coherent property to generate. *A fortiori*, no stipulational mechanism could.

In fact, even the ET, as stated, is inconsistent with the concession. For according to the ET, although there is such a property as analyticity, of necessity no sentence has it. Yet according to the concession, there could be sentences – namely, those built up in appropriate ways out of the expressions implicated in stipulative definitions – that are analytic. So even the ET needs to be modified, if it is to be made consistent with Quine's admission, thus:

> (ET*) There is a coherent property expressed by 'is analytic,' but, with the exception of those instances that are generated by stipulational mechanisms, it is necessarily uninstantiated.

Let me bring the exegetical aspect of this discussion to a premature and artificial close. It is clear that a thesis of either form would result in a philosophically important skepticism about Frege-analyticity. What we need to do is distinguish between the two theses and assess the case that can be made on their behalf.

In actual fact, however, I don't propose to look at Quine's well-known arguments in detail. Instead, my strategy will be to argue that neither a non-factualism about Frege-analyticity, nor an error thesis about it, can plausibly fall short of an outright rejection of meaning itself. Since – along with practically everybody else – I consider such a rejection to be highly implausible, I take this to constitute a *reductio* of Quine's skepticism about Frege-analyticity.

Non-factualism about Frege-Analyticity

Let's begin with the non-factualist rejection of Frege-analyticity. Now, to say that there is no such property as the property of Frege-analyticity is essentially to say that, for *any* sentence, there is no fact of the matter as to whether it is transformable into a logical truth by the substitution of synonyms for synonyms. Presumably, this itself is possible only if either there is no fact of the matter about what counts as a logical truth, or no fact of the matter about when two expressions are synonymous. Since the factuality of logic is not in dispute, the only option is a non-factualism about synonymy.

But, now, how can there fail to be facts about whether any two expressions – even where these are drawn from within a *single* speaker's idiolect – mean the same? Wouldn't this have to entail that there are no facts about what each expression means individually? Putting the question the other way: Could there be a fact of the matter about what each expression means, but no fact of the matter about whether they mean the same?[11]

Let's consider this question first against the background of an unQuinean relational construal of meaning, according to which an expression's meaning something is a relation M between it and its meaning, the meaning C. Someone who held that a non-factualism about synonymy could coexist with a determinacy about meaning would have to hold that, although it might be true that some specific word – say, "cow" – bears some specific relation M to some specific meaning C, there is no fact of the matter about whether some *other* word – some other orthographically identified particular – bears precisely the same relation to precisely the same meaning.

But how could this be? How could it conceivably turn out that it is intelligible and true to say that "cow" bears M to C, and that it is not merely false but *non-factual* to say that some other word – "vache," as it may be – also does? What could be so special about the letters "c," "o," "w"?

The answer, of course, is that there is nothing special about them. If it is factual that one word bears M to C, it is surely factual that some other word does. Especially on a relational construal of meaning, it makes no sense to suppose that a determinacy about meaning could coexist with a non-factualism about synonymy.

The question naturally arises whether this result is forthcoming *only* against the background of a relational construal of meaning. I think it's quite clear that the answer is 'no.' To see why, suppose that instead of construing meaning-facts as involving relations to meanings we construe them thus: "cow" means *cow* just in case "cow" has the monadic property R, a history of use, a disposition, or whatever your favorite candidate may be. Precisely the same arguments go through: it remains equally difficult to see how, given that "cow" has property R, it could fail to be factual whether or not some other word does.

The Error Thesis about Frege-Analyticity

I think, then, that if a plausible skepticism about Frege-analyticity is to be sustained, it cannot take the form of a non-factualism. Does an error thesis fare any better? According to this view, although there are determinate facts about which sentences are transformable into logical truths by the appropriate manipulations of synonymy, this property is necessarily uninstantiated: it is nomically impossible for there to be any Frege-analytic sentences. Our question is: Does at least this form of skepticism about Frege-analyticity avoid collapse into the indeterminacy doctrine?

Well, I suppose that if we are being very strict about it, we may have to admit that it is barely *logically possible* to combine a denial of indeterminacy with an error thesis about synonymy; so that we can say that although there are determinate facts about what means what, it is impossible for any two things to mean the same thing. But is such a view plausible? Do we have any reason for believing it? I think not.

Let's begin with the fact that even Quine has to believe that it is possible for two *tokens of the same orthographic type* to be synonymous, for that much is presupposed by his own account of logical truth. As we saw in the passage I quoted above, Quine describes a truth of logic as:

> a statement [which is true and which] remains true under all reinterpretations of its components other than the logical particles. (Quine, 1953, pp. 22–23)

Clearly, the idea isn't that such a statement will remain true no matter how the non-logical particles are substituted for, but rather that it will remain true provided that the non-logical particles are substituted for in a uniform way, with multiple occurrences of the same word receiving the same substitution in every case. But what should we count as the same here? As Strawson pointed out, it won't do merely to insist that multiple occurrences of a word be replaced by orthographically uniform replacements; for it certainly seems possible to imagine an orthographically uniform way of substituting for the non-logical particles of 'No unmarried man is married' that results in a falsehood: 'No unilluminated book is illuminated.' And it's hard to see how this is to be fixed without making some use of the idea that the orthographically uniform replacements should express the same meaning (Strawson, 1971, p. 117).

So even Quine has to admit – what in any event seems independently compelling – that two tokens of the same type can express the same meaning.

What about two tokens of different types? Here again, our own argument can proceed from Quine's own admissions. As we saw, even Quine has to concede that two expressions can mean the same thing, provided that they are explicitly stipulated to mean the same thing. So the skepticism about synonymy has to boil down to the following, somewhat peculiar claim: Although there is such a thing as the property of synonymy; and although it can be instantiated by pairs of tokens of the same orthographic type; and although it can be instantiated by pairs of tokens of distinct orthographic types, provided that they are related to each other by way of an explicit stipulation; it is, nevertheless, in principle impossible to generate instances of this property in some other way, via some other mechanism. For example, it is impossible that two expressions that were introduced independently of each other into the language should have been introduced with exactly the same meanings.

But what conceivable rationale could there be for such a claim? As far as I am able to tell, there is precisely one argument in the literature that is supposed to provide support for this claim. It may be represented as follows:

> Premise: Meaning is radically holistic in the sense that: "What our words mean depends on *everything* we believe, on *all* the assumptions we are making." (Harman, 1973, p. 14, emphasis in the original)

Therefore,

> Conclusion: It is very unlikely that, in any given language, there will be two words of distinct types that mean exactly the same thing.

I am inclined to agree that this argument (properly spelled out) is valid, and so, that if a radical holism about meaning were true, then synonymies between expressions of different types would be rare.

However, I note that "rare" does not mean the same as "impossible," which is the result we were promised. And, much more importantly, I am completely inclined to disagree that "Two dogmas" provides any sort of cogent argument for meaning holism in the first place.

It's easy to see why, if such a radical meaning holism were true, synonymies might be hard to come by. For although it is not unimaginable, it is unlikely that two words of

distinct types will participate in *all* of the same beliefs and inferences. Presumably there will always be some beliefs that will discriminate between them – beliefs about their respective shapes, for example.

But what reason do we have for believing that *all* of a word's uses are constitutive of its meaning?

Many Quineans seem to hold that the crucial argument for this intuitively implausible view is to be found in the concluding sections of "Two dogmas." In those concluding sections, Quine argues powerfully for the epistemological claim that has come to be known as the Quine–Duhem thesis: confirmation is holistic in that the warrant for any given sentence depends on the warrant for every other sentence. In those concluding sections, Quine also assumes a Verificationist theory of meaning, according to which the meaning of a sentence is fixed by its method of confirmation. Putting these two theses together, one can speedily arrive at the view that a word's meaning depends on *all* of its inferential links to other words, and hence at the thesis of meaning holism.[12]

This, however, is not a very convincing train of thought. First, and not all that importantly, this couldn't have been the argument that *Quine* intended against Frege-analyticity, for this argument for meaning holism is to be found in the very last pages of "Two dogmas," well after the rejection of Frege-analyticity is taken to have been established.

Second, and more importantly, the argument is not very compelling because it depends crucially on a verificationism about meaning, a view that we have every good reason to reject, and which has in fact been rejected by most contemporary philosophers.

Finally, and perhaps most importantly, any such holism-based argument against the possibility of synonymy would need to be supported by something that no one has ever provided – a reason for believing that yielding such an intuitively implausible result about synonymy isn't itself simply a *reductio* of meaning holism[13] (see Chapter 15, HOLISM).

III

The Analyticity of Logic

If the preceding considerations are correct, then there is no principled objection to the existence of Frege-analyticities, and, hence, no principled objection to the existence of statements that are knowable *a priori* if logical truth is.[14]

But what about logical truth? Is it knowable *a priori*? And, if so, how?[15]

In the case of some logical truths, the explanation for how we have come to know them will be clear: we will have deduced them from others. So our question concerns only the most elementary laws of sentential or first-order logic. How do we know *a priori*, for example, that all the instances of the law of non-contradiction are true, or that all the instances of *modus ponens* are valid?

As I noted above, Frege thought it obvious that there could be no substantive answer to such questions; he was inclined, therefore, to take appearances at face value and to simply *assume* the apriority of logic.

What Frege probably had in mind is the following worry. 'Explaining our knowledge of logic' presumably involves finding some *other* thing that we know, on the basis of which our knowledge of logic is to be explained. However, regardless of what that other thing is taken

to be, it's hard to see how the use of logic is to be avoided in moving from knowledge of that thing to knowledge of the relevant logical truth. And so it can come to seem as if any account of how we know logic will have to end up being vacuous, presupposing that we have the very capacity to be explained.

Michael Dummett has disputed the existence of a real problem here. As he has pointed out, the sort of circularity that's at issue isn't the gross circularity of an argument that consists of including the conclusion that's to be reached among the premises. Rather, we have an argument that purports to prove the validity of a given logical law, at least one of whose inferential steps must be taken in accordance with that law. Dummett calls this a "pragmatic" circularity. He goes on to claim that a pragmatic circularity of this sort will be damaging only to a justificatory argument that

> is addressed to someone who genuinely doubts whether the law is valid, and is intended to persuade him that it is.… If, on the other hand, it is intended to satisfy the philosopher's perplexity about our entitlement to reason in accordance with such a law, it may well do so.[16]

The question whether Dummett's distinction fully allays Frege's worry is a large one, and I can't possibly hope to settle it here. If something along these general lines can't be made to work, then *any* explanation of logic's apriority – or aposteriority, for that matter – is bound to be futile, and the Fregean attitude will have been vindicated.

However, the question that particularly interests me in the present chapter is this: Assuming that the very enterprise of explaining our knowledge of logic isn't shown to be hopeless by Frege's straightforward argument, is there any *special* reason for doubting an explanation based on the notion of analyticity? Quine's enormously influential claim was that there is. I shall try to argue that there isn't – that, in an important sense to be specified later on, our grasp of the meaning of logical claims can explain our *a priori* entitlement to holding them true (provided that the Fregean worry doesn't defeat all such explanations in the first place).

The Classical View and Implicit Definition

It's important to understand, it seems to me, that the analytic theory of the apriority of logic arose indirectly, as a by-product of the attempt to explain in what a grasp of the meaning of the logical constants consists. Alberto Coffa lays this story out very nicely in his book (Coffa, 1991, ch. 14).[17]

What account are we to give of our grasp of the logical constants, given that they are not explicitly definable in terms of *other* concepts? Had they been explicitly definable, of course, we would have been able to say, however plausibly, that we grasp them by grasping their definitions. But as practically anybody who has thought about the matter has recognized, the logical constants are not explicitly definable in terms of other concepts, and so we are barred from giving that account. The question is, what account are we to give?

Historically, many philosophers were content to suggest that the state of grasping these constants was somehow primitive, not subject to further explanation. In particular, such a grasp of the meaning of, say, 'not' was to be thought of as prior to, and independent of, a

decision on our part as to which of the various sentences involving 'not' are to count as true. We may call this view, following Wittgenstein's lead, the doctrine of

> *Flash-Grasping:* We grasp the meaning of, say, 'not' "in a flash" – prior to, and independently of, deciding which of the sentences involving 'not' are true.

On this historically influential picture, Flash-Grasping was combined with the doctrine of Intuition to generate an epistemology for logic:

> *Intuition:* Grasp of the concept of, say, negation, along with our intuition of its logical properties, explains and justifies our logical beliefs involving negation – for example, that 'If not not p, then p' is true.

As Coffa shows, this picture began to come under severe strain with the development of alternative geometries. Naturally enough, an analogous set of views had been used to explain the apriority of geometry. In particular, a flash-grasp of the indefinables of geometry, along with intuitions concerning their necessary properties, was said to explain and justify belief in the axioms of Euclidean geometry.

However, with the development of alternative geometries, such a view faced an unpleasant dilemma. Occupying one horn was the option of saying that Euclidean and non-Euclidean geometries are talking about the *same* geometrical properties, but disagreeing about what is true of them. But this option threatens the thesis of Intuition: If in fact we learn geometrical truths by intuition, how could this faculty have misled us for so long?

Occupying the other horn was the option of saying that Euclidean and non Euclidean geometries are talking about *different* geometrical properties – attaching different meanings to, say, 'distance' – and so not disagreeing after all. But this option threatens the doctrine of Flash-Grasping. Suppose we grant that a Euclidean and a non-Euclidean geometer attach different meanings to 'distance.' In what does this difference consist? Officially, of course, the view is that one primitive state constitutes grasp of Euclidean distance, and another that of non-Euclidean distance. But in the absence of some further detail about how to tell such states apart, and about the criteria that govern their attribution, this would appear to be a hopelessly *ad hoc* and non-explanatory maneuver.

The important upshot of these considerations was to make plausible the idea that grasp of the indefinables of geometry consists precisely in the adoption of one set of truths involving them, as opposed to another. Applied to the case of logic, it generates the semantical thesis that I'll call

> *Implicit Definition:* It is by arbitrarily stipulating that certain sentences of logic are to be true, or that certain inferences are to be valid, that we attach a meaning to the logical constants. More specifically, a particular constant means that logical object, if any, which would make valid a specified set of sentences and/or inferences involving it.

Wittgenstein expressed this reversal of outlook well:

> It looks as if one could *infer* from the meaning of negation that "¬¬p" means p. As if the rules for the negation sign *follow from* the nature of negation. So that in a certain sense there is first of all negation, and then the rules of grammar.

> We would like to say: "Negation has the property that when it is doubled it yields an affirmation." But the rule doesn't give a further description of negation, it constitutes negation. (Wittgenstein, 1976, pp. 52–53, cited in Coffa, 1991, p. 262)

Now, the transition from this sort of implicit definition account of grasp to the analytic theory of the apriority of logic can seem pretty immediate. For it would seem that the following sort of argument is now in place:

(1) If logical constant C is to mean what it does, then argument-form A has to be valid, for C means whatever logical object in fact makes A valid.
(2) C means what it does.

Therefore,

(3) A is valid.

I will return to various questions regarding this form of justification below.[18] For now I want to worry about the fact that neither Carnap nor Wittgenstein was content merely to replace Flash-Grasping with Implicit Definition. Typically, both writers went on to embrace some form of irrealism about logic. Intuitively, the statements of logic appear to be fully factual statements, expressing objective truths about the world, even if necessary and (on occasion) obvious ones. Both Carnap and Wittgenstein, however, seemed inclined to deny such an intuitive realism about logic, affirming in its place either the thesis of logical Non-Factualism or the thesis of logical Conventionalism, or, on occasion, both theses at once.

By logical Non-Factualism,[19] I mean the view that the sentences of logic that implicitly define the logical primitives do not express factual claims and, hence, are not capable of genuine truth or falsity. How, on such a view, are we to think of their semantic function? On the most popular version, we are to think of it as prescriptive, as a way of expressing a rule concerning the correct use of logical expressions. By contrast, logical Conventionalism is the view that, although the sentences of logic are factual – although they can express truths – their truth-values are not objective, but are, rather, determined by our conventions.

Despite this important difference between them, there is an interesting sense in which the upshot of both views is the same, a fact which probably explains why they were often used interchangeably and why they often turn up simultaneously in the analytic theory of logic. For what both views imply is that, as between two different sets of decisions regarding which sentences of logic to hold true, there can be no epistemic fact of the matter. In short, both views imply an epistemic relativism about logic. Conventionalism implies this because it says that the truth in logic is up to us, so no substantive disagreement is possible; and Non-Factualism implies this because it says that there are no truths in logic, hence nothing to disagree about.

Nevertheless, for all this affinity of upshot, it should be quite plain that the two views are very different from – indeed, incompatible with – each other. Conventionalism is a factualist view: it presupposes that the sentences of logic have truth-values. It differs from a realist view of logic in its conception of the *source* of those truth-values, not on their existence. Therefore, although it is possible, as I have noted, to find texts in which

a rule-prescriptivism about logic is combined with Conventionalism, that can only be a confusion.

The important question is: Why did the proponents of Implicit Definition feel the need to go beyond it all the way to the far more radical doctrines of logical Non-Factualism and/or Conventionalism? Whatever problems it may eventually be discovered to harbor, Implicit Definition seems like a plausible candidate for explaining our grasp of the logical constants, especially in view of the difficulties encountered by its classical rival. But there would appear to be little that *prima facie* recommends either logical Non-Factualism or logical Conventionalism. So why combine these dubious doctrines with what looks to be a plausible theory of meaning?

Apparently, both Carnap and Wittgenstein seem to have thought that the issue was forced, that Implicit Definition logically entailed one or the other anti-realist thesis. It seems quite clear that Carnap, for example, believed that Implicit Definition brought Conventionalism immediately in its wake; and Quine seems to have agreed. What separated them was their attitude towards Conventionalism. Carnap embraced it; Quine, by contrast, seems to have been prepared to reject any premise that led to it, hence his assault on the doctrine of Implicit Definition.

But if this is in fact the correct account of Quine's motivations, then they are based, I believe, on a false assumption, for neither form of irrealism about logic follows from the thesis of Implicit Definition.

I will proceed as follows. First, I will argue that Implicit Definition, properly understood, is completely independent of any form of irrealism about logic. Second, I will defend the thesis of Implicit Definition against Quine's criticisms. Finally, I will examine the sort of account of the apriority of logic that this doctrine is able to provide.

Implicit Definition and Non-Factualism

Does Implicit Definition entail Non-Factualism? It is certainly very common to come across the claim that it does. Coffa, for instance, writes that from the new perspective afforded by the doctrine of Implicit Definition, the basic claims of logic are

> our access to certain meanings, definitions in disguise, devices that allow us to implement an explicit or tacit decision to constitute certain concepts.… From this standpoint, necessary claims do not tell us anything that is the case both in the world and in many others, as Leibniz thought, or anything that is the case for *formal* reasons, whatever that might mean, or anything that one is forced to believe due to features of our mind. They do not tell us anything that is the case; so they had better not be called claims or propositions. Since their role is to constitute meanings and since (apparently) we are free to endorse them or not, it is better to abandon the old terminology (a priori "principles," "laws," etc.) that misleadingly suggests a propositional status and to refer to them as "rules." (Coffa, 1991, pp. 265–266)

I have no desire to engage the exegetical issues here; as far as I can tell, the middle Wittgenstein seems very much to have been a non-factualist about the implicit definers of logic, just as Coffa says. What I dispute is that it *follows* from the fact that a given sentence Q is being used to implicitly define one of its ingredient terms, that Q is not a factual sentence, not a sentence that "tells us anything that is the case." These two claims seem to me to be entirely independent of each other.

To help us think about this, consider Kripke's example of the introduction of the term 'meter.' As Kripke imagines it, someone introduces the term into his vocabulary by stipulating that the following sentence is to be true:

(1) Stick S is a meter long at t.

Suppose that stick S exists and is a certain length at t. Then it follows that 'meter' names that length and hence that (1) says that stick S is that length at t, and since it is that length at t, (1) is true.

Knowing all this may not be much of an epistemic achievement, but that isn't the point. The point is that there appears to be no inconsistency whatsoever between claiming that a given sentence serves to implicitly define an ingredient term and claiming that that very sentence expresses something factual.

Similarly, I don't see that there is any inconsistency between supposing that a given logical principle – for instance, the Law of Excluded Middle – serves to implicitly define an ingredient logical constant, and supposing that that very sentence expresses a factual statement capable of genuine truth and falsity.[20]

Implicit Definition and Conventionalism

So far I have argued that it is consistent with a sentence's serving as an implicit definer that that very sentence come to express a fully factual claim, capable of genuine truth and falsity. Perhaps, however, when implicit definition is at issue, the truth of the claim that is thereby fixed has to be thought of as conventionally determined? Does at least Conventionalism follow from Implicit Definition?[21]

It is easy to see, I suppose, why these two ideas might have been run together. For according to Implicit Definition, 'if, then,' for example, comes to mean the conditional precisely by my assigning the truth-value True to certain basic sentences involving it; for example, to

If, if p then q, and p, then q.

And in an important sense, my assigning this sentence the value True is arbitrary. Prior to my assigning it that truth-value, it didn't have a complete meaning, for one of its ingredient terms didn't have a meaning at all. The process of assigning it the value True is simply part of what fixes its meaning. Had I assigned it the value False, the sentence would then have had a *different* meaning. So, prior to the assignment there couldn't have been a substantive question regarding its truth-value. And after the assignment there couldn't be a substantive question as to whether that assignment was correct. In this sense, then, the sentence's truth-value is arbitrary and conventional. Doesn't it follow that Implicit Definition entails Conventionalism?

Not at all. All that is involved in the thesis of Implicit Definition is the claim that the conventional assignment of truth to a sentence determines what proposition that sentence expresses (if any); such a view is entirely silent about what (if anything) determines the truth of the claim that is thereby expressed – *a fortiori*, it is silent about whether our conventions determine it.

Think here again of Kripke's meter stick. If the stick exists and has such-and-so length at t, then it is conventional that 'meter' names that length and, therefore, conventional that (1) expresses the proposition *Stick S has such-and-so length at t*. However, that stick S has that length at t is hardly a fact generated by convention; it presumably had that length prior to the convention, and may continue to have it well after the convention has lapsed.[22]

I anticipate the complaint that the entailment between Implicit Definition and Conventionalism is blocked only through the tacit use of a distinction between a sentence and the proposition it expresses, a distinction that neither Carnap nor Quine would have approved.

Such a complaint would be mistaken, however. The argument I gave relies not so much on a distinction between a sentence and a proposition in the technical sense disapproved of by Quine, as on a distinction between a sentence and *what it expresses*. And it is hard to see how any adequate philosophy of language is to get by without some such distinction.[23] Even on a deflationary view of truth, there is presumably a distinction between the *sentence* 'Snow is white' and that which makes the sentence true, namely, snow's being white. And the essential point for my purposes is that it is one thing to say that 'Snow is white' comes to express the claim that snow is white as a result of being conventionally assigned the truth-value True; and quite another to say that snow comes to be white as a result of our conventions. The first claim is Implicit Definition (however implausibly applied in this case); and the other is Conventionalism. Neither one seems to me to entail the other.

Quine against Implicit Definition: Regress

As I noted above, I am inclined to believe that erroneous opinion on this score has played an enormous role in the history of this subject. I conjecture that had Quine felt more confident that Implicit Definition could be sharply distinguished from Conventionalism, he might not have felt so strongly against it.

In any event, though, whatever the correct explanation of Quine's animus, we are indebted to him for a series of powerful critiques of the thesis of Implicit Definition, critiques that have persuaded many that that thesis, and with it any explanation of the apriority of logic that it might be able to ground, are fundamentally flawed. We must now confront Quine's arguments.

According to Implicit Definition, the logical constants come to have a particular meaning in our vocabulary by our conventionally stipulating that certain sentences (or inferences) involving them are to be true. For instance, let us assume that the meaning for 'and' is fixed by our stipulating that the following inferences involving it are to be valid:

$$(2)\quad \frac{\text{A and B}}{\text{A}} \qquad \frac{\text{A and B}}{\text{B}} \qquad \frac{\text{A, B}}{\text{A and B}}$$

Now, Quine's first important criticism of this idea occurs in his early paper "Truth by convention" (1976a).[24] As Quine there pointed out, there are an infinite number of instances of schema (2). Consequently, the inferences of this infinitary collection could not have

been conventionally stipulated to be valid singly, one by one. Rather, Quine argued, if there is anything at all to this idea, it must be something along the following lines: We adopt certain general conventions, from which it follows that all the sentences of the infinitary collection are assigned the value Valid. Such a general convention would presumably look like this:

> Let all results of putting a statement for 'p' and a statement for 'q' in 'p and q implies p' be valid.

However, the trouble is that in order to state such a general convention we have had, unavoidably, to use all sorts of logical terms – 'every,' 'and,' and so on. So the claim, essential to the proposal under consideration, that all our logical constants acquire their meaning via the adoption of such explicitly formulated conventional assignments of validity must fail. Logical constants whose meaning is not fixed in this way are presupposed by the model itself.[25]

This argument of Quine's has been very influential; and I think that there is no doubt that it works against its target as specified. However, it is arguable that its target as specified isn't the view that needs defeating.

For, surely, it isn't compulsory to think of someone's following a rule R with respect to an expression e as consisting in his *explicitly stating* that rule in so many words in the way that Quine's argument presupposes. On the contrary, it seems far more plausible to construe x's following rule R with respect to e as consisting in some sort of fact about x's *behavior* with e.

In what would such a fact consist? Here there are at least a couple of options. According to a currently popular idea, following rule R with respect to e may consist in our being disposed to conform to rule R in our employment of e, under certain circumstances. On this version, the notion of rule-following would have been *reduced* to a certain sort of dispositional fact. Alternatively, one might wish to appeal to the notion of following a given rule, while resisting the claim that it can be reduced to a set of naturalistically acceptable dispositional facts. On such a non-reductionist version there would be facts about what rule one is following, even if these are not cashable into facts about one's behavioral dispositions, however optimal (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §2).

For myself, I am inclined to think that the reductionist version won't work, that we will have to employ the notion of following a rule unreduced.[26] But because it is more familiar, and because nothing substantive hangs on it in the present context, I will work with the reductionist version of rule-following. Applied to the case we are considering, it issues in what is widely known in the literature as a "conceptual role semantics" (CRS).

According to this view, then, the logical constants mean what they do by virtue of figuring in certain inferences and/or sentences involving them, and not in others. If some expressions mean what they do by virtue of figuring in certain inferences and sentences, then some inferences and sentences are *constitutive* of an expression's meaning what it does, and others aren't. And any CRS must find a systematic way of saying which are which, of answering the question: What properties must an inference or sentence involving a constant C have, if that inference or sentence is to be constitutive of C's meaning?

Quine against Implicit Definition: Constitutive Truth

Now, Quine's second objection to Implicit Definition can be put by saying that there will be no way of doing what I said any CRS must do – namely, systematically specify the meaning-constituting inferences. Quine formulated this point in a number of places. Here is a version that appears in "Carnap and logical truth":

> if we try to warp the linguistic doctrine of logical truth into something like an experimental thesis, perhaps a first approximation will run thus: *Deductively irresoluble disagreement as to a logical truth is evidence of deviation in usage (or meanings) of words*.… [However] the obviousness or potential obviousness of elementary logic can be seen to present an insuperable obstacle to our assigning any experimental meaning to the linguistic doctrine of elementary logical truth.… For, that theory now seems to imply nothing that is not already implied by the fact that elementary logic is obvious or can be resolved into obvious steps. (Quine, 1976b, p. 105)

Elsewhere, Quine explained his use of the word "obvious" in this connection thus:

> In "Carnap and Logical Truth" I claimed that Carnap's arguments for the linguistic doctrine of logical truth boiled down to saying no more than that they were obvious, or potentially obvious – that is, generable from obvieties by obvious steps. I had been at pains to select the word 'obvious' from the vernacular, intending it as I did in the vernacular sense. A sentence is obvious if (a) it is true and (b) any speaker of the language is prepared, for any reason or none, to assent to it without hesitation, unless put off by being asked so obvious a question. (Quine, 1975, p. 206)

Quine's important point here is that there will be no substantive way of distinguishing between a highly obvious, non-defining sentence and a sentence that is an implicit definer. Both types of sentence – if, in fact, both types exist – will have the feature that any speaker of the language will be prepared to assent to instances of them, "for any reason or none." So in what does the alleged difference between them consist? How is distinctive content to be given to the doctrine of Implicit Definition?[27]

Now, there is no doubt that this is a very good question; and the belief that it has no good answer has contributed greatly to the rejection of the doctrine of Implicit Definition. Fodor and Lepore, for example, base the entirety of their recent argument against a conceptual role semantics on their assumption that Quine showed this question to be unanswerable (Fodor and Lepore, 1991a).

If Quine's challenge is allowed to remain unanswered, then the threat to the analytic theory of the *a priori* is fairly straightforward. For if there is no fact of the matter as to whether S is a sentence that I must hold true if S is to mean what it does, then there is no basis on which to argue that I am entitled to hold S true without evidence.

But that would seem to be the least of our troubles, if Quine's argument is allowed to stand; for what's threatened is not only the apriority of logical truths but, far more extremely, the *determinacy* of what they claim. For as I've already pointed out, and as many philosophers are anyway inclined to believe, a conceptual role semantics seems to be the *only* plausible view about how the meaning of the logical constants is fixed. It follows, therefore, that if there is no fact of the matter as to which of the various inferences involving a constant are meaning-constituting, then there is also no fact of the matter as to what the logical

constants themselves mean. And that is just the dreaded indeterminacy of meaning on which the critique of analyticity was supposed not to depend.

The simple point here is that if the only view available about how the logical constants acquire their meaning is in terms of the inferences and/or sentences that they participate in, then any indeterminacy in what those meaning-constituting sentences and inferences are will translate into an indeterminacy about the meanings of the expressions themselves. This realization should give pause to any philosopher who thinks he can buy in on Quine's critique of Implicit Definition without following him all the way to the far headier doctrine of meaning-indeterminacy.

There has been a curious tendency to miss this relatively simple point. Fodor seems a particularly puzzling case; for he holds all three of the following views. (1) He rejects indeterminacy, arguing forcefully against it. (2) He follows Quine in rejecting the notion of a meaning-constituting inference. (3) He holds a conceptual role view of the meanings of the logical constants. As far as I am able to judge, however, this combination of views is not consistent.[28]

Part of the explanation for this curious blindness derives from a tendency to view Quine's argument as issuing not in an indeterminacy about meaning, but, rather, in a *holism* about it. In fact, according to Fodor and Lepore, the master argument for meaning holism in the literature runs as follows:

(1) Some of an expression's inferential liaisons are relevant to fixing its meaning.
(2) There is no principled distinction between those inferential liaisons that are constitutive and those that aren't. (The Quinean result)

Therefore,

(3) All of an expression's inferential liaisons are relevant to fixing its meaning. (Meaning Holism)

Fearing this argument's validity, and seeing no way to answer Quine's challenge, Fodor and Lepore spend their whole book trying to undermine the argument's first premise, namely, the very plausible claim that at least *some* of an expression's inferential liaisons are relevant to fixing its meaning (see Fodor and Lepore, 1991b).

But they needn't have bothered, for I don't see how the master argument could be valid in the first place. The claim that *all* of an expression's inferential liaisons are constitutive of it cannot cogently follow from the claim that it is *indeterminate* what the constitutive inferences are. If it's *indeterminate* what the constitutive inferences are, then it's genuinely *unsettled* what they are. And that is inconsistent with saying that they are *all* constitutive, and inconsistent with saying that *none* are constitutive, and inconsistent with saying that some specified subset are constitutive.

Fodor and Lepore are not alone in not seeing the problem here. Let me cite just one more example. In his comments on an earlier version of the present chapter, Harman says:

> Can one accept Quine's argument against analyticity without being committed to the indeterminacy of meaning? Yes and no. By the "indeterminacy of meaning" might be meant an indeterminacy as to which of the principles one accepts determine the meanings of one's terms and

> which simply reflect one's opinions about the facts. Clearly, Quine's argument against analyticity is committed to that sort of indeterminacy. [However] that by itself does not imply full indeterminacy in the sense of Chapter 2 of *Word and Object*. (Harman, 1994b)

As Harman correctly says, Quine has to deny that there is a fact of the matter as to which of T's principles determine the meanings of his terms and which simply reflect T's opinions about the facts – that, after all, is just what it is to deny that there are facts about constitutivity. However, Harman insists, this denial in no way leads to the indeterminacy thesis of chapter 2 of *Word and Object*.

But this is very puzzling. Against the background of a conceptual role semantics, according to which the meaning of T's term C is determined precisely by a certain subset of the principles involving C that T accepts, an indeterminacy in what the meaning-determining principles are will automatically lead to an indeterminacy in what the meaning is, in the full sense of chapter 2 of *Word and Object*. If a subset (not necessarily proper) of accepted principles is supposed to determine meaning; and if there is no fact of the matter as to which subset that is; then there is, to that extent, no fact of the matter as to what meaning has been determined. Since correct translation is supposed to preserve meaning, it follows that there can be no fact of the matter as to what counts as correct translation.

I think there is really no avoiding the severe conclusion that meaning is indeterminate, if the Quinean challenge to constitutivity is allowed to remain unanswered. I'm inclined to think, therefore, that anyone who rejects radical indeterminacy of meaning must believe that a distinction between the meaning-constituting and the non-meaning-constituting can be drawn. The only question is how.

Well, that is not the task of the present chapter. Although there are some good ideas about this, I don't have a fully thought-through proposal to present just now.[29] My main aim here is not to *solve* the fundamental problem for a conceptual role semantics for the logical constants; rather, as I have stressed, it is to show that, against the background of a rejection of indeterminacy, its insolubility cannot be conceded.

Pending the discovery of other problems, then, it seems open to us to suppose that a plausible theory of meaning for the logical constants is given by something like the following:

> A logical constant C expresses that logical object, if any, that makes valid its meaning-constituting inferences or sentences.

Implicit Definition, Justification, and Entitlement

Now, how does any of this help vindicate the analytic theory of the apriority of logic, the idea that logic is epistemically analytic? Let us consider a particular inference form, A, in a particular thinker's (T) repertoire; and let's suppose that that inference form is constitutive of the meaning of one of its ingredient constants C. How, exactly, might these facts help explain the epistemic analyticity of A for T?

To say that A is epistemically analytic for T is to say that T's knowledge of A's meaning alone suffices for T's justification for A, so that empirical support is not required. And it does seem that a conceptual role semantics can provide us with a model of how

that might be so. For given the relevant facts, we would appear to be able to argue as follows:

(1) If C is to mean what it does, then A has to be valid, for C means whatever logical object in fact makes A valid.
(2) C means what it does.

Therefore,

(3) A is valid.

Now, it is true that this is tantamount to a fairly broad use of the phrase "knowledge of the meaning of A," for this knowledge includes not merely knowledge of what A means, strictly so called, but also knowledge of how that meaning is fixed. But this is, of course, both predictable and unavoidable: there was never any real prospect of explaining apriority merely on the basis of a knowledge of propositional content. Even Carnap realized that one needed to know that a given inference or sentence had the status of a 'meaning postulate.'

But isn't it required, if this account is to genuinely explain T's *a priori* justification for the basic truths of logic, that T know the premises *a priori* as well? Yet, it hasn't been shown that T can know the premises *a priori*.

It is quite correct that I have not attempted to show that the relevant facts about meaning cited in the premises are knowable *a priori*, although I believe that it is intuitively quite clear that they are. I have purposely avoided discussing all issues relating to knowledge of meaning-facts. My brief here has been to defend epistemic analyticity; and this requires showing only that certain sentences are such that, *if* someone knows the relevant facts about their meaning, *then* that person will be in a position to form a justified belief about their truth. It does not require showing that the knowledge of those meaning-facts is itself *a priori* (although, I repeat, it seems quite clear to me that it will be).[30]

Isn't it a problem for the aspirations of the present account that a thinker would have to use *modus ponens* to get from the premises to the desired conclusion?

Not if Dummett's distinction between pragmatic and vicious circularity is credited with opening a space for an epistemology for logic, as discussed above.

Finally, how could such an account possibly hope to explain the man in the street's justification for believing in the truths of logic? For such a person, not only would the relevant meaning-facts be quite opaque, he probably wouldn't even be capable of framing them. Yet such a person is obviously quite justified in believing the elementary truths of logic. Thus, so our objector might continue, this sort of account cannot explain our ordinary warrant for believing in logic; at best, it can explain the warrant that sophisticates have.

I think that, strictly speaking, this objection is correct, but only in a sense that strips it of real bite. Philosophers are often in the position of articulating a warrant for an ordinary belief that the man in the street would not understand. If we insist that a person counts as justified only if they are aware of the reason that warrants their belief, then we will simply have to find another term for the kind of warrant that ordinary folk often have and that philosophers seek to articulate. Tyler Burge has called it an "entitlement":

> The distinction between justification and entitlement is this. Although both have positive force in rationally supporting a propositional attitude or cognitive practice, and in constituting an

> epistemic right to it, entitlements are epistemic rights or warrants that need not be understood by or even be accessible to the subject.… The unsophisticated are entitled to rely on their perceptual beliefs. Philosophers may articulate these entitlements. But being entitled does not require being able to justify reliance on these resources, or even to conceive such a justification. Justifications, in the narrow sense, involve reasons that people have and have access to. (Burge, 1993, p. 458)

When someone is entitled, all the facts relevant to the person's justification are already in place, so to say; what's missing is the reflection that would reveal them.

Just so in the case at hand. If a conceptual role semantics is true, and if A is indeed constitutive of C's meaning what it does, then those facts by themselves constitute a warrant for A; empirical support is not necessary. A can only be false by meaning something other than what it means. But these facts need not be known by the ordinary person. They suffice for his entitlement, even if not for his full-blown justification. This full-blown justification can be had only by knowing the relevant facts about meaning.

Conclusion

Quine helped us see the vacuity of the metaphysical concept of analyticity and, with it, the futility of the project it was supposed to underwrite – the linguistic theory of necessity. But I don't see that those arguments affect the epistemic notion of analyticity that is needed for the purposes of the theory of *a priori* knowledge. Indeed, it seems to me that epistemic analyticity can be defended quite vigorously, especially against the background of a realism about meaning.

On the assumption that our warrant for believing in elementary logical truths cannot be explained, the outstanding problem is to explain our *a priori* knowledge of conceptual truths. For this purpose, the crucial semantical notion is that of Frege-analyticity. I have argued that this notion is bound to be in good standing for a meaning realist.

If the project of explaining logic is not ruled hopeless, then I have tried to show how the doctrine that appears to offer the most promising account of how we grasp the meanings of the logical constants – namely, Implicit Definition – can explain the epistemic analyticity of our logical beliefs and, hence, our *a priori* warrant for believing them. As long as we are not prepared to countenance radical indeterminacy, we should have every confidence that this form of explanation can be made to work.

Appendix: *A Priori* Knowledge of the Second Premise

I have argued that a conceptual role semantics supplies the following sort of warrant for our belief in the elementary truths of logic.

(1) If C is to mean what it does, then A has to be valid, for C means whatever logical object in fact makes A valid.
(2) C means what it does.

Therefore,

(3) A is valid.

In this Appendix I want to propose a reason for holding that the second premise in this argument form is knowable *a priori*.

The challenge might appear, at first, to be utterly trivial. Surely, we know, for any given C, that it means whatever it means. Suppose C is the word "and"; then, surely, we know *a priori* that "and" means whatever it means. Indeed, isn't it clear that we know *a priori* precisely what it does mean, namely, *and*? For any given mentioned constant, isn't disquotation guaranteed to state its meaning accurately?

What all such purely disquotational views of our knowledge of meaning ignore is the possibility that the words we are disquoting fail to have a meaning in the first place. What the disquotational maneuver guarantees is only that, *if* a word has a meaning, then disquotation will state its meaning correctly. However, the disquotational view does not, and cannot, address the question of how we know that the word has a meaning to begin with.

This point is interestingly related to a point made by Harman in the following passage:

> Even if conventional assignments of truth or falsity determine meaning, it does not follow that a sentence is true by virtue of convention. It does not even follow that the sentence is true. (Harman, 1968, pp. 130–131; see also Quine, 1976a, pp. 93–95; 1976b, p. 114)

Harman's claim is that, even if we put aside objections to the thesis of Implicit Definition, it wouldn't follow that a meaning-constituting sentence is true. Hence, we couldn't claim to be entitled to S without evidence, just because S is meaning-constituting. Perhaps S is meaning-constituting and not true.

How might this happen? Harman doesn't explain: but it's important to ask. How might it turn out that a sentence that is stipulated to be true, as a way of fixing the meaning of some ingredient term t, nevertheless fail to be true?

One thing is, I think, certain: not by being false. For to be false, S would have to be meaningful. And it is stipulated that, if S expresses any meaning at all, it expresses a true one. Under these assumptions, therefore, S can fail to be true only by expressing no meaning whatever. And this in turn will happen only if one of its ingredient terms fails to express a meaning.

So let us ask: How might it turn out that a set of constitutive rules for a term t fail to determine a meaning for it? I can think of two ways. First, the meaning-constituting role specified for t may impose inconsistent demands on it, thus making it impossible for there to be a meaning that makes true all of its meaning-constituting sentences. A second worry might arise simply against the background of a robust propositionalism, without exploiting worries about inconsistency. For according to a robust propositionalism, meanings are radically mind-independent entities whose existence no amount of defining could ensure. Hence, there may well not be a meaning answering to all the demands placed upon a term by a set of stipulations.[31]

For both of these reasons, then, we cannot immediately conclude from the fact that t is governed by a set of meaning-constituting rules, that t is meaningful.

To put this point in terms of the second premise of the argument-form outlined above, the fact that a given constant C is governed by certain constitutive rules of use doesn't by itself entitle us to conclude that C means what it does, because it doesn't by itself entitle us to conclude that C has a meaning in the first place. Hence, we cannot lean on a disquotational view to vindicate our claim that the second premise is knowable *a priori*.

A Solution

So how are we to proceed? Is there an *a priori* way of laying to rest a doubt about the meaningfulness of our logical constants? I think we can make a case for the following claim: we are *a priori* entitled to believe that our basic logical constants are meaningful because we cannot coherently doubt that they are. For the assumption that our constants are meaningful is presupposed in any attempt to claim that they aren't.

I shall assume that, prior to having seen the desirability of introducing an alternative set of logical constants, we start off with a particular, single set of these. For the sake of specificity, let us assume that these constants are classical, that is, that they are governed by classical constitutive rules. As will become clear, this assumption is absolutely inessential to the argument that follows.

Now, consider what a first attempt to formulate a skepticism about the meaningfulness of our constants would look like. The Skeptic wishes to assert that our basic logical constants do not express a meaning. Given the assumption that our constants are classical, the Skeptic's assertion comes down to a claim about the meaningfulness of the basic pair of constants in terms of which all the others can be defined – let's suppose that that pair consists of negation and the conditional. In the case of negation, the Skeptic would appear to want to claim:

(4) ∀x (If x is a token of 'not,' then x does not have a meaning);

and in the case of 'if, then' this:

(5) ∀x (If x is a token of 'if, then,' then x does not have a meaning).

The problems with both (4) and (5), however, are not easy to miss. In attempting to state that our basic logical constants fail to be meaningful, both claims have to assume that those very constants *are* meaningful. No one could rationally wish to assert (4) or (5) who did not believe that negation and the conditional are meaningful; yet what (4) and (5) claim is that negation and the conditional are not meaningful (respectively). It would appear, therefore, that any attempt to assert (4) and (5) would be self-defeating: the very act of putting those propositions forward undermines the truth of the propositions that are being put forward.

A number of possible lines of objection need to be considered. First, does the argument especially depend on the particular selection of basic logical constants? Would it work equally well if, say, negation and disjunction had been chosen as the reduction base, rather than negation and the conditional?

It is hard to see how the particular selection can make any difference. It is relatively trivial to show that, regardless of the particular choice of reduction base, the same style of argument goes through, with similar effect.

A second, somewhat more challenging line of objection runs as follows. Let's concede that an assertion of (4) would be self-defeating: we cannot use the constants we have to assert of *them* that they are meaningless.[32] But why couldn't we introduce some *new* constants and use them to formulate a skepticism about the old ones? (Notice that I don't need to make any assumptions about what these prior constants are, whether classical or otherwise.) Such a claim would look like this:

(6) ∀x (If x is a token of 'not,' then x does not_{N} refer),

where the subscript 'N' indicates that the constant is one of the *new* ones. There appears to be nothing self-stultifying about this thesis.

I don't think that this objection works either, though its problems are slightly better hidden. The problem is that the integrity of the old constants is presupposed in *the very act of introducing the new alternatives*.

Recall, we are operating on the assumption that we start off with a determinate set of logical constants. By further, and ultimately optional, assumption, these constants are classical. Now we wish to introduce an alternative set of constants, so that we may use them to state, of the old constants, that they are meaningless. Consider how such an introduction would have to go. We would need to say what the constitutive rules governing the new constants are, and that they are *all* of them. That is, we would need to say something along the following lines (where the subscript 'N' indicates that the constant is *new*):

(7) ∀x (If x is a token of 'not_N,' then x is subject to rules R1, R2, R3, and no others).

Clearly, however, in this definition the meaningfulness of many of the old constants – in particular, negation, the conditional, and the universal quantifier – is presupposed. And there appears to be no way to cancel that presupposition without jeopardizing the meaningfulness of the new constants with the use of which the skeptical hypothesis is to be formulated. Unless the old constants *are* meaningful, the stipulations will fail to give the new constants a particular meaning. Hence, we cannot coherently suppose both that the new constants have a meaning and that the old ones don't.

As far as I am able to judge, every attempt to formulate a worry about the meaningfulness of our basic logical constants runs into a similar sort of difficulty: every such attempt ends up presupposing the integrity of the constants whose integrity it seeks to question.

An enormous number of questions are left outstanding, none of which can be adequately dealt with here. For one: Is this merely a pragmatic result, or something stronger? Tentative answer: Something stronger. To sustain the claim that the result is merely pragmatic, one would have to make sense of the claim that, although we cannot rationally doubt that our constants are meaningful, it is nevertheless possible that they aren't. However, considerations similar to the ones adduced above would tend to show that we cannot make sense of this thought either.

Another question: Doesn't this argument prove too much? Some of the best recent philosophy has taken the form of claiming that various of the rules of classical logic make unsatisfiable demands on our ability to mean what we do by our words and, hence, that they are incoherent. Isn't any criticism of this form disabled by the above argument?

Not at all. My argument does nothing to preclude the following sort of view: Our constants are essentially intuitionistic – that is, they are governed by a core set of intuitionistic rules. However, some philosophers and mathematicians have mistakenly supposed that they are also subject to certain further rules – they have mistakenly supposed, in other words, that our constants are classical. However, they are mistaken in this: not only are our constants not classical, but they couldn't have been, because creatures like us are incapable of meaning classical constants.

Nothing in my argument prevents someone from adopting the sort of view outlined in the preceding paragraph. What my argument does preclude is the simultaneous assertion that our ordinary constants are classical *and* that they are incoherent. As far as I can see,

though, no one with an interest in criticizing classical logic need put his position in that manifestly problematic way.[33,34]

Notes

1 Consider, for example, the following passage from Quine (1970, p. 3):

> My objection to recognizing propositions does not arise primarily from philosophical parsimony – from a desire to dream of no more things in heaven and earth than need be. Nor does it arise, more specifically, from particularism – from a disapproval of intangible or abstract entities. My objection is more urgent. If there were propositions, they would induce a certain relation of synonymy or equivalence between sentences themselves: those sentences would be equivalent that expressed the same proposition. Now my objection is going to be that the appropriate equivalence relation makes no objective sense at the level of sentences. This, if I succeed in making it plain, should spike the hypothesis of propositions.

2 As I say, I am going to work with this linguistic picture out of deference to my opponents. I would prefer to work with a propositionalist picture of belief. Most of the crucial notions developed in this chapter, and much of the argument involving them, can be translated, with suitable modifications, into this propositionalist framework. Thus, even those who believe, as I do, that knowledge is not a matter of knowing that certain sentences are true can find use for the account developed here.

3 The inclusion of the word "outer" here is partly stipulative. I have always found it natural to regard *a priori* knowledge as encompassing knowledge that is based on no experience as well as knowledge that is based purely on *inner* experience.

4 In the interests of brevity, I shall henceforth take it as understood that "justification" means "justification with a strength sufficient for knowledge."

5 Even this strong notion is not as demanding as many have supposed. For instance, it is consistent with a belief's being *a priori* in the strong sense that we should have *pragmatic* reasons for dropping it from our best overall theory. For illuminating discussion of the modesty of the notion of the *a priori*, see Wright (1984) and Hale (1986, ch. 6).

6 I am indebted to Paul Horwich for emphasizing the importance of this point.

7 See Frege (1950). (Some may regard the attribution of precisely this notion to Frege controversial. What matters to me is not who came up with the idea, but rather the philosophical role it has played.)

My use of the term 'analytic' in connection with Frege's *semantical* notion as well as with the preceding epistemic and metaphysical concepts may be thought ill advised. But I do so deliberately, to highlight the fact that the term has been used in the literature in general, and in Quine in particular, to stand for all three different sorts of notion, often without any acknowledgement of that fact. This terminological promiscuity has undoubtedly contributed to the confusion surrounding discussions of this issue.

8 For some discussion, see Boghossian (1994a).

9 Exegetically, this does leave us with a couple of puzzles. First, "Two dogmas" does contain a brief discussion of the implicit definition idea, under the guise of the notion of a "semantical rule." Given that, why does Quine insist that he intends only to discuss the notion of Frege-analyticity? Second, the notion of a semantical rule is discussed only in connection with non-logical truths: since, however, the deployment of this idea would be exactly the same in the logical case, why is the analyticity of logic expressly excluded? Third, given that the analyticity of logic is expressly excluded, on what basis does Quine allow himself to draw morals about logic's revisability towards the end of "Two dogmas"? I think there is no avoiding the conclusion that, on this and other related issues (see below), "Two dogmas" is confused. It would, in fact, have been surprising if these rather tricky problems had all been in clear focus in Quine's pioneering papers.

10 In this context, nothing fancy is meant by the use of such expressions as 'property' and 'proposition.' For present purposes they may be understood in a thoroughly deflationary manner.

I have sometimes been asked why I consider just this particular weakening of a non-factualist thesis, one that involves, problematically from Quine's official point of view, a modal notion? Why not rather attribute to him the following *Very Weak Thesis*:

> (VWT) There is a coherent, determinate property expressed by 'is analytic,' but *as a matter of fact*, it has never been instantiated; consequently, all tokens of the sentence 'S is analytic' have been false up to now.

There are two reasons. First, the VWT is not a philosophically interesting thesis; and, second, it could not have been argued for on the basis of a *philosophy* paper – i.e., on the sorts of *a priori* grounds that Quine offers. So although Quine may not be entitled to precisely the ET, I am going to ignore that and not hold it against him.

11 This question was first asked by Grice and Strawson (1989). Grice and Strawson didn't sufficiently stress, however, that Quine was committed to a skepticism even about *intralinguistic* synonymy, and not just about inter-linguistic synonymy, for the theory of apriority doesn't much care about the inter-linguistic case.

12 Formulations of this argument may be found in Fodor (1987, pp. 62 ff.), Fodor and Lepore (1991b, pp. 37 ff.), and Devitt (1995, p. 17). None of the authors mentioned approve of the argument.

13 A further "Two dogmas"-based argument for meaning holism, this time invalid, will be considered further below, in connection with the discussion of the thesis of Implicit Definition.

14 As before, subject to the proviso about the apriority of synonymy.

15 I am ignoring for now the class of *a priori* truths that are neither logical nor Frege-analytic. As we shall see, the very same strategy – implicit definition – that can be applied to explain our knowledge of logic can be applied to them as well.

16 Dummett (1991, p. 202). Dummett's distinction is deployed in a somewhat different context.

17 In the next three paragraphs I follow the general contours of the account that Coffa develops. However, the formulations are mine and they differ in important respects from Coffa's, as we shall see further on.

18 Readers who are acquainted with a paper of mine entitled "Inferential role semantics and the analytic/synthetic distinction" (Boghossian, 1994b), will be aware that I used to worry that Implicit Definition could not generate *a priori* knowledge because of the falsity of something I called "The Principle." The Principle is the thesis that it follows from a sentence's being an implicit definer that that sentence is true. The proper place of this issue in the overall dialectic, and a proposed solution, are discussed in the Appendix to the present chapter.

19 Not to be confused with the non-factualism about Frege-analyticity discussed earlier in the chapter.

20 Someone may object that the two cases are not relevantly analogous. For the meter case is supposed to be a case of the *fixation of reference*, but the logical case an instance of the fixation of meaning. Doesn't this difference between them block the argument I gave?

I don't see that it does. First, the two cases really are disanalogous only if there is an important difference between meaning and reference; yet, as is well known, there are many philosophers of language who are inclined to think that there isn't any such important difference. Second, it seems to me that even if we allowed for a robust distinction between meaning and reference, the point would remain entirely unaffected. Whether we think of an implicit definer as fixing a term's reference directly, or as first fixing its meaning, which then in turn fixes its reference, seems to me entirely irrelevant to the claim that Implicit Definition does not entail Non-Factualism. As long as both processes are consistent with the fixation of a factual claim for the sentence at issue – as they very much seem to be – the point stands.

21 Certainly many philosophers seem to have thought so. Richard Creath, for example, sympathetically expounds Carnap's view that the basic axioms of logic implicitly define the ingredient

logical terms by saying that on this view "the postulates (together with the other conventions) create the truths that they, the postulates express." See Creath (1992, p. 147).

22 This point is also forcefully made by Salmon (1993) and Yablo (1992).

23 Notice that conventionalists themselves need to make crucial use of such a distinction when they describe their own position, as in the passage cited above from Creath: "the postulates (together with the other conventions) create the truths that they, the postulates, express." As Hilary Putnam pointed out some time ago, it's hard to see how distinctive content is to be given to Conventionalism without the use of some such distinction. For a conventionalism merely about linguistic expressions is trivial. A real issue is joined only when the view is formulated as a claim about the truths expressed. See Putnam (1975).

24 Quine's argument here is officially directed against a Conventionalism about logical truth, that is, against the idea that logical truth is determined by our conventions. This idea we have already rejected in our discussion of the metaphysical concept of analyticity. However, Quine attacks Conventionalism *by* attacking the semantical thesis of Implicit Definition. Hence the need for the present discussion.

25 Quine claims that this argument may also be put as follows: The claim that the sentences of logic lack assignment of truth-value until they are conventionally assigned such values must fail. For logic is needed in order to infer from a formulated general convention that the infinitely many instances of a given schema are true. Hence, sentences of logic whose truth-value is not fixed as the model requires are presupposed by the model itself.

It's unclear to me that this is a formulation of precisely the same argument. However, to the extent that it is distinct, it is also addressed by the proposal I put forth below.

26 For discussion, see Boghossian (1989).

27 For all its influence, it is still possible to find the force of the Quinean point being underestimated by the friends of Implicit Definition. Christopher Peacocke, for example, in a subtle defense of an inferential role semantics, claims that what makes the inferences involving the logical constants constitutive is that a thinker finds those inferences "primitively compelling," and does so because they are of those forms. He goes on to explain:

> To say that a thinker finds such instances primitively compelling is to say this: (1) he finds them compelling; (2) he does not find them compelling because he has inferred them from other premises and/or principles; and (3) for possession of the concept in question … he does not need to take the correctness of the transitions as answerable to anything else. (Peacocke, 1992, p. 6)

I think it is plain, however, that these conditions are insufficient for answering the Quinean challenge: a non-constitutive, though highly obvious, form of inference may also be found compelling because of its form, and not on the basis of inference from anything else. So these conditions cannot be what distinguish between a constitutive and a non-constitutive inference.

28 For Fodor's views on the mentioned issues, see Fodor (1987; 1994).

29 For a good start, see Peacocke (1992).

30 For a discussion of why the second premise is *a priori* see the Appendix to the present chapter.

31 This, I believe, was the basis of Arthur Prior's worry about an inferential role semantics; it was unfortunate that he tried to illustrate his point in a way that misleadingly suggested that his was a worry of the first sort, about consistency. See Prior (1967).

32 This was suggested to me in conversation by Hartry Field.

33 For much more on all this, see my "Knowledge of logic" (2000).

34 I am grateful to a number of audiences – at MIT, CUNY Graduate Center, Michigan State, the University of Chicago, the SOFIA Conference on Tenerife, the Chapel Hill Colloquium, Dartmouth College, London University, and Oxford University. An earlier version of this chapter was presented at the NEH Institute on the "Nature of Meaning," held at Rutgers University in the summer of 1993. It was there that I first became aware that Christopher Peacocke has been thinking

along somewhat similar lines about the *a priori* – see his "How are a priori truths possible?" (1993) presented at the Rutgers conference. Although there are a number of differences between our approaches, and although Peacocke's focus is not on the notion of analyticity, I have benefited from discussing these matters with him. Another philosopher to whom I am grateful for numerous illuminating conversations is Jerry Katz. Although Katz carves up the issues in this area very differently than I do, he deserves an enormous amount of credit for keeping the topic of analyticity alive during a period when it was extremely unfashionable to do so. I also benefited from presenting a version of this chapter as part of a symposium on Analytic Truth, involving Gil Harman, Burton Dreben, and W.V.O. Quine, at the 1994 Eastern Division meetings of the APA. I am especially grateful to Gil Harman, Elizabeth Fricker, Hartry Field, Gary Gates, Bill Lycan, Stephen Schiffer, and Barry Loewer for their detailed comments on previous versions of this chapter. Special thanks are due to Bob Hale and Crispin Wright for their patience and for their very helpful reactions to several different drafts. For other helpful discussion and commentary, I want to thank Jennifer Church, Jerry Fodor, Albert Casullo, Norma Yunez, Neil Tennant, Peter Unger, Tom Nagel, Paul Horwich, Ned Block, Richard Creath, Allan Gibbard, Stephen Yablo, and David Velleman.

References

Boghossian, P. 1989. "The rule-following considerations." *Mind*, 98(392): 507–549.

Boghossian, P. 1994a. "The transparency of mental content." *Philosophical Perspectives*, 8: 33–50.

Boghossian, P. 1994b. "Inferential role semantics and the analytic/synthetic distinction." *Philosophical Studies*, 73(2–3): 109–122.

Boghossian, P. 2000. "Knowledge of logic." In *New Essays on the A Priori*, edited by P. Boghossian and C. Peacocke, pp. 229–254. Oxford: Oxford University Press.

Burge, T. 1993. "Content preservation." *Philosophical Review*, 102(4): 457–488.

Coffa, A. 1991. *The Semantic Tradition*. Cambridge: Cambridge University Press.

Creath, R. 1992. "Carnap's conventionalism." *Synthese*, 93(1–2): 141–165.

Devitt, M. 1995. *Coming to our Senses*. New York: Cambridge University Press.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. 1994. *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, J., and E. Lepore. 1991a. "Why meaning (probably) isn't conceptual role." *Mind & Language*, 6(4): 328–343.

Fodor, J., and E. Lepore. 1991b. *Holism: A Shopper's Guide*. Oxford: Blackwell.

Frege, G. 1950. *The Foundations of Arithmetic*. Translated by J. L. Austin. Oxford: Blackwell.

Grice, H. P., and P. Strawson. 1989. "In defense of a dogma." In *Studies in the Way of Words*, edited by H. P. Grice, pp. 196–212. Cambridge, MA: Harvard University Press.

Hale, B. 1986. *Abstract Objects*. Oxford: Blackwell.

Harman, G. 1968. "Quine on meaning and existence, I: the death of meaning." *Review of Metaphysics*, 21(1): 124–151.

Harman, G. 1973. *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. 1994a. "Doubts about conceptual analysis." In *Philosophy in Mind*, edited by M. Michael and J. O'Leary-Hawthorne, pp. 43–48. Dordrecht, Netherlands: Kluwer.

Harman, G. 1994b. "Comments on Boghossian." APA Symposium on Analytic Truth, Boston, MA.

Lycan, W. 1991. "Definition in a Quinean world." In *Definitions and Definability: Philosophical Perspectives*, edited by J. Fetzer, D. Shatz, and G. Schlesinger, pp. 111–131. Dordrecht, Netherlands: Kluwer.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. 1993. "How are a priori truths possible?" *European Journal of Philosophy*, 1(2): 175–199.

Prior, A. 1967. "The runabout inference ticket." Reprinted in Strawson, 1967, pp. 129–131.

Putnam, H. 1975. "The refutation of conventionalism." In *Mind, Language and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.
Quine, W. V. O. 1953. "Two dogmas of empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1960. *Word and Object*. Boston: MIT Press.
Quine, W. V. O. 1970. *The Philosophy of Logic*. Englewood Cliffs, NJ: Prentice Hall.
Quine, W. V. O. 1975. "Reply to Hellman." In *The Philosophy of W. V. O. Quine*, edited by P. A. Schilpp, p. 206. La Salle: Open Court.
Quine, W. V. O. 1976a. "Truth by convention." Reprinted in *The Ways of Paradox*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1976b. "Carnap and logical truth." Reprinted in *The Ways of Paradox*. Cambridge, MA: Harvard University Press.
Salmon, N. 1993. "Analyticity and apriority." *Philosophical Perspectives*, 7: 125–133.
Strawson, P., ed. 1967. *Philosophical Logic*. Oxford: Oxford University Press.
Strawson, P. 1971. *Logico-Linguistic Papers*. London: Methuen.
Wittgenstein, L. 1976. *Philosophical Grammar*. Los Angeles: University of California Press.
Wright, C. 1984. "Inventing logical necessity." In *Language, Mind and Logic*, edited by J. Butterfield, pp. 187–209. Cambridge: Cambridge University Press.
Yablo, S. 1992. "Review of Sidelle." *Philosophical Review*, 101: 878–881.

Further Reading

Carnap, R. 1947. *Meaning and Necessity*. Chicago: University of Chicago Press.
Dummett, M. 1973. *Frege: The Philosophy of Language*. London: Duckworth.
Dummett, M. 1978. *Truth and Other Enigmas*. London: Duckworth.
Dummett, M. 1991. *Frege: The Philosophy of Mathematics*. Cambridge, MA: Harvard University Press.
Field, H. 1977. "Logic, meaning and conceptual role." *Journal of Philosophy*, 74(7): 379–409.
Pap, A. 1958. *Semantics and Necessary Truth*. New Haven: Yale University Press.
Putnam, H. 1975. *Mind, Language and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.
Putnam, H. 1979. "Philosophy of logic." Reprinted in *Mathematics, Matter and Method: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.
Wright, C. 1980. *Wittgenstein on the Foundations of Mathematics*. Cambridge, MA: Harvard University Press.

Postscript: Further Thoughts about Analyticity: 20 Years Later

PAUL ARTIN BOGHOSSIAN

Introduction

One central thesis of the chapter[1] to which the present piece is an afterword was that two fundamentally distinct notions had been conflated under the label 'analytic' – an epistemological notion and a metaphysical one.[2]

A second central thesis was that while the critics of analyticity, Quine (1953) and Harman (1967) principally among them, were right to disparage the metaphysical notion, the epistemological notion could be shown to be cogent and to serve a useful role in the theory of *a priori* justification.

According to the metaphysical notion,

> A sentence S is *metaphysically analytic* if and only if it is true (false) by virtue of its meaning alone (without any contribution from the world).

By contrast, the epistemic notion has it that

> A sentence S is *epistemically analytic* if and only if it is possible to justifiably believe S merely by virtue of understanding S's meaning (and without any contribution from sensory experience).

There had been a pervasive tendency in the literature to conflate these two ideas, as is illustrated by the following representative passage from BonJour (1998, p. 28):

> the moderate empiricist position on a priori knowledge holds that while such knowledge genuinely exists … it is nonetheless merely analytic in character – that is, very roughly, merely a product of human concepts, meaning, definitions, or linguistic conventions. Such knowledge thus says nothing substantive about the world.

In the 20 or so years since "Analyticity" was first published, the importance of distinguishing between the metaphysical and epistemological concepts has come to be widely accepted. It has become commonplace in discussions of analyticity to cite the distinction and to respect its substance.[3]

More controversial than the distinction itself has proven to be my claim that epistemic analyticity can play a significant role in explaining *a priori* justification. Also contentious, although to a much lesser degree, has been my rejection of the notion of metaphysical analyticity.

Gillian Russell (2008) put up an interesting defense of the metaphysical notion. Her arresting thought is that developments in our understanding of the notion of meaning, since Quine's famous discussion, provide previously unavailable routes to a coherent conception of 'truth solely in virtue of meaning.'

In my (2011) I explain why I am not persuaded by Russell's arguments; but I believe, nonetheless, that they deserve the considerable attention they have received. Recently, Bob Hale and Crispin Wright (forthcoming) and Jared Warren (2014) have also advanced important defenses of the metaphysical notion. David Liggins (ms.) has an interesting analysis of some of the framework issues involved.

A more vigorous debate has centered on the claim that epistemological analyticity can play a central role in the epistemology of *a priori* justification. In connection with this issue, four big questions arise:

1. Can we explain *all a priori* justification via epistemic analyticity?
2. Assuming not, does that imply that we can't explain *any a priori* justification through epistemic analyticity? Is uniformity a requirement on the explanation of justification in the *a priori* domain?

3. Assuming uniformity is not a requirement, is there *any a priori* justification that we can explain via epistemic analyticity?
4. If we can explain some, how would it work?

I will look, inevitably too briefly, at some developments in my thinking about these issues since the initial appearance of "Analyticity."

Can Epistemic Analyticity Explain All *A Priori* Justification?

Throughout my explorations of the idea of epistemic analyticity I've never claimed that it can explain *all* cases of *a priori* justification. I was mostly concerned to show that the epistemically analytic was capable of explaining our knowledge of conceptual truths and of the basic truths and inferences of logic.

However, any proponent of epistemic analyticity must eventually confront the question whether all *a priori* justification can be explained in these terms. Over the years, it has become clear to me that the answer to this question has to be 'No.'

The famous case of color exclusion, for example, resists explanation in terms of epistemic analyticity. I know *a priori* that nothing can be red and green all over at the same time. But knowledge of this fact is not encoded in my grasp of the relevant color concepts, as can be seen from the following considerations. Knowledge of the exclusion does not reside in my grasp of the concept *red,* because I could have the concept *red* without so much as having the concept *green.* The concept *red* couldn't be speaking negatively about its compatibility with green, so to say, because it doesn't speak about it at all. Vice versa for the concept *green*.

At best, then, knowledge of the exclusion would have to reside in the *joint possession* of *red* and *green*. But isn't the joint possession of *red* and *green* just the simultaneous possession of each? If that's right, then joint possession can't contain any information that's not contained in the sum of what's contained in each.

As a result, we can't explain our *a priori* knowledge of the way in which red excludes green merely on the basis of our understanding of the ingredient color concepts. If we are nevertheless able to figure it out we must be relying on something else.[4]

A second kind of case that's problematic for epistemic analyticity arises when the concepts in question *do* speak of a necessary relation obtaining between two properties but where that doesn't suffice for claiming that the relation obtains in reality.

This seems to be the case with *normative* truths. Take morality. On the basis of a Trolley thought experiment, you might conclude that it is not morally permissible to throw the fat man off the bridge in order to stop the oncoming trolley and save the five innocent persons strapped to the rails below. But it would not be plausible to claim that your judgment derives solely from your understanding of the ingredient concepts.

The reason why is related to Moore's Open Question Argument (Moore, 1903). If someone tells us that, according to his concept of *good*, the good always involves some particular property – maximizing happiness, for example – we can always ask: But is that the *correct* concept of good, the one that delivers genuinely normative results, as opposed to simply telling us what's good according to your concept?

Notice the contrast here with the concept *square*. If someone said: According to my concept *square*, a square always has four sides, it wouldn't make any sense to ask: But is that the

correct concept *square*? By contrast, it always seems to make sense to ask this question about correctness of any particular normative concept.

If this is right, then we can't explain our *a priori* moral knowledge on the basis of our understanding of the ingredient concepts. If we are nevertheless able to figure it out, it seems as though we must be relying on something else.

Is Uniformity a Requirement?

If the argument up to this point is correct, then not all *a priori* justification can be explained via the understanding alone. But is it open to us to suppose that justification in the *a priori* domain may have more than one source? Doesn't the argument up to this point imply that *no a priori* justification can be explained via epistemic analyticity?

I don't believe that all instances of *a priori* justification need have the same source.

There is an interesting contrast here with the empirical. Because the 'empirical' has a positive characterization, there is no question that it has a unified source – namely, sensory experience. (That's not to deny, of course, that there are different subspecies of empirical knowledge – for example, inferential versus observational.)

But the *a priori* is defined as that which is *not* empirical. And this means that it is not guaranteed up front that everything that is *a priori* will have the same epistemic basis. For all that the notion of the *a priori* implies it's possible that some *a priori* justification has its source in understanding and the rest in some other source, for example, intuition.

We should be open to the suggestion that *a priori* justification is generated in several different ways, using different resources.

Can Epistemic Analyticity Explain Even Some Cases of *A Priori* Justification?

But can we in fact explain *any* cases of *a priori* justification through epistemic analyticity?

Let us look at sentences and inferences where such explanations seem most plausible:

(1) All squares have four sides.
(2) All foxes are foxes.
(3) If all vixens are foxes, and all foxes are mammals, then all vixens are mammals.
(4) All vixens are foxes and all foxes are mammals, so: All foxes are mammals.
(5) Anyone who knows p believes p.

If a sentence S is epistemically analytic, we have said, then an understanding of its meaning suffices for our being able to justifiably assent to it. But suffices how?

This fundamental issue was not clear in "Analyticity." I got a lot clearer about it in my (2003b) and subsequently. Once the issue is clarified, it becomes much harder to see exactly how explanations in terms of epistemic analyticity are supposed to work, even while it continues to look extremely plausible that at least some of our *a priori* knowledge should be explained in this way.

Broadly speaking, there are two very different routes from grasp to justifiable assent, depending on how we think about the relation between grasp and assent. On one way of thinking about it,

(Constitutive)
Grasp of S's meaning is in part *constituted* by a disposition to assent to S.

On an alternative way of thinking about it,

(Basis)
Grasp of S's meaning is constituted by something distinct from the disposition to assent to S but provides a potential *epistemic basis* for the disposition to assent.

The former, *Constitutive*, option in effect construes the understanding of a given sentence in *conceptual role* terms. To grasp S's meaning is to be disposed to assent to S under certain conditions, or to be prepared to make certain inferences involving S.

For example, one influential view has it that for you to mean *conjunction* by 'and' you must be prepared to infer according to (4) and in general from any sentence of the form 'A and B' to 'A.' Similarly, to understand (1) you must be prepared to assent to it.

On the alternative Basis view, grasp of p is distinct from any disposition to assent to p, and so can serve, and sometimes does serve, as the epistemic basis for assenting to p.

The Constitutive View

Let me begin with the first option, the Constitutive view. One limitation of such a view is that a conceptual role semantics, viewed simply as a theory of meaning, has always seemed most plausible in the case of the logical constants, and perhaps also of theoretical terms, but not so much in application to other concepts. If this is right, then the Constitutive view will have limited application.

Williamson (2003; 2007; 2012) has denied that it is plausible as a theory of meaning in even the most favorable cases. He denies that there are any constitutive understanding-assent links.

I don't believe that Williamson's case against the very existence of such constitutive understanding-assent links succeeds; but I won't argue for that here (see my 2012).

However, I do agree that a conceptual role semantics is not plausible for many types of concept – in particular, for color concepts or moral concepts.

More importantly for our purposes, even in those cases where a disposition to assent is plausible as a theory of grasp, it's not at all clear why the grasp-constituting dispositions come out *justified* as a mere consequence of the fact that they are grasp-constituting.

It is tempting to think otherwise (a fact to which I can attest). Suppose that my possessing the concept *and* is constituted in part by my disposition to infer from 'A and B' to 'A.' It's tempting to think that I'm justified in making such an inference merely because it is constitutive of my possession of the concept *and*.

But how is that supposed to work? Is any disposition that is built into the possession of a given concept thereby justified? My (2003a) presented a series of counter-examples to such a generalized meaning-justification connection, and proposed a much more restricted

principle bridging concept grasp and justification. A number of critics (see, for example, Schechter and Enoch, 2006) have highlighted various problems for my proposal.

But the consideration that turned me decisively against Constitutive accounts was one that I had been dimly aware of all along, but which I hadn't confronted properly, and it's this: Even if we got a bridge principle that returned the right verdicts on particular cases, the resultant account of epistemic justification would still be irretrievably *externalist*, since it won't in general be introspectively accessible to the subject which of his dispositions is concept constituting. And I reject externalist accounts of epistemic justification on the grounds that they distort the essentially normative character of the notion of justification.[5]

To summarize: There are three large problems for a Constitutive version of an epistemic analyticity account of *a priori* justification.

- Conceptual role theories are plausible only for a limited range of concepts.
- It's not clear that there is a plausible meaning-justification connection that is extensionally correct.
- Any such connection looks to deliver only an externalist justification for the disposition to assent, one that would be opaque to the subject.

The Basis View

Do *Basis* accounts fare any better?

On a Basis account, the understanding of p is constituted somehow or other, the crucial point being that, since it is not constituted by facts involving assent to p, it can serve as an *epistemic basis* for assent to p.

Of course, we have neither a settled view of understanding, nor a settled view of what an epistemic basis is – and these facts make it hard to flesh out such an account in the requisite detail. We can be confident, though, that both of these notions are in good standing, being needed quite generally, beyond the context of our immediate concerns.

What we are not entitled to be confident about, however, is that when we finally do get satisfying accounts of grasp and basis, it will be clear how grasp of a proposition p could serve as an *epistemic basis* for assent to p.

When I base my belief that the cat is on the mat on my experience of the cat's being on the mat, my grasp enables me to think the relevant thought assent to which I then base on my sensory experience. But what would it be for the grasp itself to be my basis?

There is one instance of this sort of basing that we may be said to understand reasonably well. And that is when my grasp of p consists in my grasp of some sort of explicit definition for p; and my basis for assenting to p consists in my inferring p from its definition (Frege-analyticity).

But this is clearly a very special case – special both in that grasp rarely consists in grasping an *explicit definition*, and in that basis rarely means *inferential basis*.

But if the relation between grasp and assent is not like that, what else could it be like?

The only other model that we have for something's serving as an epistemic basis for a belief is that which obtains between the perceptual state that p and assent to p. But how could the relation between grasp and assent be analogous to the relation that obtains in this case?

A perceptual state can serve as an epistemic basis for assent because it is a *presentation* of the world as being a certain way, a *seeming* – that's essential to its ability to rationalize belief.[6]

But it is hard to see how the grasp of the meaning of a sentence could be a presentation of anything.

Setting this problem to one side, there is another difficulty that any Basis account must confront. If the relation between understanding and the disposition to assent is contingent, as the Basis account would have it, then it is presumably *possible* for someone to understand p perfectly well and yet, even after extended reflection, refuse to assent to p. For example, most people who understand

(5) Anyone who knows p believes p

assent to it. However, some experts on knowledge, who understand (5) perfectly well, refuse to assent to it. They assent to its negation (Williamson's example). On the Constitutivist account, according to which assent is constitutive of understanding, this refusal would impugn the claim that these experts understand (5) perfectly well. But no such conclusion follows on a Basis account.

How, though, on the Basis account, are we to explain why those who affirm (5) are justified on the basis of their understanding of (5), whereas those who deny (5) are not? By hypothesis, both groups understand (5) perfectly well.[7,8]

Conclusion

There has always seemed something right about epistemic analyticity – intuitively, some *a priori* justification derives from our competence with the relevant concepts or meanings.

But it has become increasingly clear that not all *a priori* justification can be explained in this way. The domain of the normative poses an especially important challenge.

Furthermore, even in those cases where an explanation in terms of the epistemically analytic seems most promising, the exact mechanism by which justification for assent may be generated by the understanding alone seems obscure and ill understood.

All of this has made me take the classical project of explaining the *a priori* in part by appeal to the rationalist notion of 'intuition' much more seriously than I had previously been inclined to do (see Boghossian, 2016; forthcoming). Needless to say, much difficult work lies ahead.

Notes

1 Boghossian (1997). A somewhat abbreviated version of the paper appeared as Boghossian (1996).

2 For the most part, I will take sentences to be the bearers of analytic truth or falsity. But everything I say can be modified easily to apply to propositions. Furthermore, I will assume a justification-first epistemology and a broadly internalist view of justification; and I will focus on doxastic justification as opposed to propositional justification.

3 For example, the distinction plays a major role in Williamson (2007) and Russell (2008); also of interest are Margolis and Laurence (2001), Horwich (2000), and Glüer (2003).

4 As I explain later on, I am now inclined to look with favor on the suggestion that the missing ingredient is intuition.

5 If I were happy with externalist conceptions of justification, I would be a reliabilist. And in that case, it would be relatively easy to see what to say about the problem of *a priori* justification.

6 Bengson (2015) has emphasized the terminology of 'presentation.' Unlike him, though, I think of presentations as equivalent to seemings.
7 This argument originates with Ernest Sosa (2007). However, Sosa doesn't embed his argument, as I think is required for it to make sense, in the framework provided by the distinction between Constitutive and Basis accounts.
8 Here, again, it can seem tempting to think that intuition supplies the missing ingredient.

References

Bengson, J. 2015. "The intellectual given." *Mind*, 124(495): 707–760.
Boghossian, P. 1996. "Analyticity reconsidered." *Noûs*, 30(3): 360–391.
Boghossian, P. 1997. "Analyticity." In *A Companion to the Philosophy of Language*, edited by B. Hale and C. Wright, pp. 331–368. Oxford: Blackwell.
Boghossian, P. 2003a. "Blind reasoning." *Aristotelian Society,* suppl. vol. 77(1): 225–248.
Boghossian, P. 2003b. "Epistemic analyticity: a defense." *Grazer Philosophische Studien*, 66(1): 15–35.
Boghossian, P. 2011. "Truth in virtue of meaning." *Australasian Journal of Philosophy*, 89(2): 370–374.
Boghossian, P. 2012. "Inferentialism and the epistemology of logic: reflections on Casalegno and Williamson." *Dialectica*, 66(2): 221–236.
Boghossian, P. 2016. "Intuitions and the understanding." In *The Present and Future of Virtue Epistemology*, edited by M. Fernandez. Oxford: Oxford University Press.
Boghossian, P. Forthcoming. "Intuitions and philosophy." In *Debating the A Priori and the Analytic*, by P. Boghossian and T. Williamson. Oxford: Oxford University Press.
BonJour, L. 1998. *In Defense of Pure Reason: A Rationalist Account of A Priori Justification*. Cambridge: Cambridge University Press.
Glüer, K. 2003. "Analyticity and implicit definition." *Grazer Philosophische Studien*, 66(1): 37–60.
Harman, G. 1967. "Quine on meaning and existence, I: the death of meaning." *The Review of Metaphysics*, 21(1): 124–151.
Hale, B., and C. Wright. Forthcoming. "Bolzano's definition of analytic propositions." In *Festschrift for Peter Simons*, edited by S. Lapointe. *Grazer Philosophische Studien*.
Horwich, P. 2000. "Stipulation, meaning, and apriority." In *New Essays on the A Priori*, edited by P. Boghossian and C. Peacocke, pp. 150–69. Oxford: Oxford University Press.
Liggins, D. (ms.) "Grounding and metaphysical analyticity."
Margolis, E., and S. Laurence. 2001. "Boghossian on analyticity." *Analysis*, 61(4): 293–302.
Moore, G. E. 1903. *Principia Ethica*. London: Dover Publications.
Quine, W. V. O. 1953. "Two dogmas of empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
Russell, G. 2008. *Truth in Virtue of Meaning: A Defence of the Analytic/Synthetic Distinction*. Oxford: Oxford University Press.
Schechter, J., and D. Enoch. 2006. "Meaning and justification: the case of modus ponens." *Noûs*, 40(4): 687–715.
Sosa, E. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford: Oxford University Press.
Warren, J. 2014. "The possibility of truth by convention." *The Philosophical Quarterly*, 65(258): 84–93.
Williamson, T. 2003. "Understanding and inference." *Aristotelian Society,* suppl. vol. 77(1): 249–93.
Williamson, T. 2007. *The Philosophy of Philosophy*. Chichester: Wiley-Blackwell.
Williamson, T. 2012. "Boghossian and Casalegno on understanding and inference." *Dialectica*, 66(2): 237–247.

24

Rule-Following, Objectivity, and Meaning

BOB HALE

1 Wittgenstein on Meaning, Understanding, and Rules

There is widespread agreement that Wittgenstein advances, in the rule-following sections of *Philosophical Investigations* and *Remarks on the Foundations of Mathematics* (Wittgenstein, 1967, §§138–242; 1978, part VI), considerations that are quite destructive of certain conceptions of meaning, understanding, and rule-following into which we may easily slide when we attempt a general philosophical account of them: that meaning something by a certain expression is a special act or state of mind, accompanying or lying behind writing or speaking; that understanding an expression consists in supplying or adopting an interpretation for it; that following a rule – a rule for the use of a word, say – is a matter of traveling along rails which are already laid down and determine its application in new cases, and so on. And it is equally generally agreed that Wittgenstein's aims, in his discussions of these matters, are not wholly negative and destructive – that he seeks to replace these misconceptions by a better account, armed with which we shall be able to resist the pressures which push us into them: using an expression according to a rule is not founded upon reasons, but that does not mean that there can be no going right (or wrong) in our use of expressions – and the key to understanding how this can be so lies in the idea that to employ an expression with a certain meaning, or according to a rule, is to participate in a custom or practice. It is, in other words, no part of his overall purpose to uphold blanket skeptical conclusions, to the effect that there are no such things as meaning something by a particular expression, as understanding another's words, or as employing an expression according to a rule (or as following a rule of any kind). His aim, rather, seems clearly enough to have been to rid us of badly mistaken pictures of what these things are, and to point us towards a proper, less inflated, conception of them.

This much is, I believe, quite uncontroversial. What is controversial is the *extent* of the destruction wrought by the negative considerations Wittgenstein advances and, consequentially, the exact character of the conception of meaning, understanding,

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

and rule-following – centered on the somewhat elusive ideas of custom and practice – that we may retain in the light of a proper appreciation of their destructive effect. There is, in particular, a sharp opposition between what may be termed 'conservative' readings, which see Wittgenstein as solely concerned to undermine certain seductive misconceptions (McDowell, 1984; Baker and Hacker, 1984a; 1984b; 1985) and count it an error to interpret him as providing support for any skeptical or revisionary theses about meaning and related matters, and more radical ones (Kripke, 1982; Wright, 1980; 1981; 1984b; Carruthers, 1984) which claim to find in his writings grounds for calling into question, in one way or another, what may roughly and provisionally be called the objectivity of meaning.

This exegetical issue will not be pursued here.[1] Even if the conservatives are right, the more skeptical lines of thought which Wittgenstein's discussions have suggested to some thinkers quite certainly merit careful attention in their own right, whether or not they can defensibly be attributed to Wittgenstein, or regarded as drawing out consequences of claims to which he uncontroversially commits himself. It is with two of these more skeptical directions of theorizing that we shall be concerned.

2 Kripke on Rules

Kripke (1982) interprets central sections of *Philosophical Investigations* (§§138–242) as developing a 'skeptical paradox' about meaning. The paradoxical conclusion of the skeptical argument is that there is *no fact about what anyone means by any expression* she uses. Faced with this seemingly outrageous conclusion, we naturally incline to the view that there *must* be something wrong with the argument leading to it: that it relies on some assumption which we can reject, or that it makes some fallacious step. To attempt to sustain this claim is to go for a 'straight solution,' which enables us to maintain that there is, after all, some species of fact in which our meaning what we do by our words consists. But Kripke argues – and takes Wittgenstein to have argued – that there can be no such meaning-constitutive facts: the argument, to be reviewed shortly, proceeds by elimination, that is, it considers the various types of fact that might be supposed to play this constitutive role, and tries to show that they cannot do the job required of them. So Kripke advocates instead a 'skeptical solution,' that is, a response to the paradox which *accepts* the skeptical conclusion but seeks to explain how we can live with it; in particular, how we can rehabilitate talk of meaning without supposing that there are facts in virtue of which meaning ascriptions (such as statements of the form 'S means such-and-such by E') are true or false.

Kripke develops the skeptical argument in terms of one central example. Suppose '$68 + 57 = ?$' is a question I have never explicitly considered. What answer should I give? I shall almost certainly answer '125.' And I shall naturally suppose that this is not only the arithmetically correct answer, but the one I must give, if my answer is to be in accord with what I have all along meant by '+' or 'plus.' It is, I suppose, a fact that when I used '+' before, I meant a certain definite function – one which has, *inter alia,* the value 125 for the arguments 68, 57, and not some other function, which has a different value for those arguments. In particular it is a fact, surely, that I didn't mean the function Kripke calls 'quus' (for which we shall use the symbol '$\oplus$'), where $m \oplus n = m + n$, provided that $m, n < 57$, but in case m or $n \geq 57$, $m \oplus n = 5$.

Kripke's skeptic maintains that there is no such fact. His argument focuses initially on the claim that '125' is the answer I must give, if I am to be in accord with what I formerly

meant by '+.' Granting, *pro tem*, that there is no problem about my *present* understanding of '+' – that I use it to mean addition – the skeptic presses two questions, one constitutive and one epistemological:

> What makes it the case that up to now I have meant addition rather than, say, quaddition by '+,' so that '125' is the answer I should give to '68 + 57 = ?,' if I am to be in agreement with what I meant by my previous uses of '+'?
> What justifies me in thinking that this is the answer I ought to return, if I am to be in agreement with my past meaning for '+'?

It is crucially important to note that these questions are posed against the background of an idealizing assumption about my cognitive powers. I am assumed to have perfect recall of all potentially relevant aspects of my past linguistic and non-linguistic behavior, and of all my preceding mental life, any previous thoughts, imaginings, or the like, which may have accompanied my previous uses of '+.' It should be noted also that there is to be no prior restriction upon the type of fact that may be admissibly cited as constitutive of meaning; in particular, there is no Quinean restriction to purely physical or behavioral facts. This idealization sets Kripke's skeptic apart from the traditional variety of epistemological skeptic: if, even under the idealizing assumption, it proves impossible to justify the claim that I meant addition, the conclusion to be drawn is not – with the traditional skeptic – that whilst there may be a determinate fact about what I meant, it lies beyond our epistemic reach, but that there simply is no such fact at all (see Kripke, 1982, pp. 14–15; Wright, 1984a, pp. 762–763; Boghossian, 1989, p. 515).

The skeptic's answer to both questions is, of course: 'Nothing.' By hypothesis, I have never confronted this particular addition problem before, so that the answer I should now give is not settled by my having previously had the explicit thought, or forming the explicit intention, to answer this question by '125.' Furthermore, my past applications of '+' are finite in number, and it is clearly consistent with my past answers to questions of the form 'm + n = ?' that I meant some other function by '+' (such as $\oplus$), which coincides with addition over the cases actually encountered, but diverges from it over '68 + 57 = ?' No finite selection of answers determines to within uniqueness what rule (if any) I was following. The skeptic then argues that no state or event in consciousness – no previous thoughts or imaginings, nor even a special experience of meaning – can constitute the needed fact. First, it is obviously questionable whether there is in fact any single conscious state or event which invariably accompanied my previous uses. Second – and more important – even if there had been, this would be powerless to settle the question unless that state or event in consciousness were itself insusceptible of alternative, quus-like interpretations. In particular, if any past state of consciousness is to prescribe answers in particular as-yet-unencountered cases, it would have to possess a *general* content – a distinctive feeling or mental picture won't do, because it will never be transparent what that requires of me in new cases; rather, it would have to be something like a general thought, such as that the answer I should give to any question of the type 'm + n = ?' is the one which I obtain by counting a collection of m marbles, say, and then a disjoint collection of n marbles, and finally counting the union of these two collections. But this gets us nowhere, unless we assume that there is no parallel problem about what I meant by the terms in which the general rule was formulated. We are just assuming that by 'count' I formerly meant what I now mean by that word, and did not mean *quount*, where quounting the union of two sets gives the same

answer as counting them, provided that neither of the sub-collections has more than 56 elements; otherwise, the result of the quount is to be '5.' Keeping this perverse interpretation in play may need further perverse hypotheses about what I meant by such terms as 'union,' 'subset,' or 'co-extensive.' But there is no evident reason to think they can't be conjured up, with a little ingenuity.[2]

In effect, Kripke's point here is that sooner or later we are going to have to deal with the situation where I am supposed to have attached a certain definite meaning to certain words without giving myself an explanation of them, or rules for applying them, in general terms – so we may as well just suppose that '+' is such.

The thought is tempting that our failure to locate a meaning-constituting fact in the details of my past applications of '+,' or in the conscious states or events which may be supposed to have accompanied them, results from our looking in the wrong place, for a fact of the wrong sort. My meaning one thing rather than another by my words consists, it may be supposed, in my being *disposed* to apply them in certain ways and not in others. The attraction of this suggestion is that it can be a perfectly good fact that I was disposed to do certain things, not others, even though I did not actually do them – for the circumstances appropriate to exercise of the disposition need not have presented themselves. In particular, it could be that I was all along disposed to answer '125' to the question '57+ 68 = ?' but never actually did so, simply because no events occurred to trigger my additive disposition in this particular way. But the dispositional proposal must be rejected, Kripke argues, for two reasons. First, although it may at first appear that linguistic dispositions have the requisite generality, this is an illusion. There are potentially infinitely many questions of the form 'm + n = ?' but it just isn't true – or so Kripke claims – that for each and every one of them I was disposed to give a certain definite answer. We can only speak correctly of my being disposed to answer this way rather than that, when the numbers to be added are not too big for me to add. In this sense, our dispositions are *finite*. But this means that the dispositional 'solution' doesn't overcome the problem about the finiteness of actual past uses, for the class of answers I did give *or would have given* is still finite; and the skeptic can then undercut the proposed solution by choosing his example so that it lies beyond the reach of my additive (quadditive?) dispositions (Kripke, 1982, pp. 26–28). Second, the dispositional proposal fails to capture the essentially *normative* aspect of meaning. I may well be disposed to make certain sorts of mistake when doing addition. If what I meant by '+' is identified with what I was disposed to say, in answer to '+' questions, then there is no room for a needed contrast between the answers I *would* have given and those which I *should* have given, the latter being those which accord with my past meaning for '+.' Generally, the claim that some expression means such-and-such has a normative component – it is a claim about the circumstances in which it is, or would be, correct to apply it – which evades capture by an attempted reduction of (putative) meaning-constitutive facts to dispositional facts (Kripke, 1982, pp. 28–32).

Taking it that the alternatives considered and rejected exhaust the possibilities, Kripke's skeptic concludes that there is no fact constitutive of my *having meant* + rather than ⊕ by '+' in the past, and nothing that could justify my conviction that '125' is the answer I should now give to '68 + 57 = ?' if I am to be faithful to my past meaning for '+.' Furthermore, the conclusion appears to admit of straightforward generalization: if there were a fact in virtue of which I *now* mean + by '+,' then – under the idealizing assumption of perfect recall, and so on – I would be able to cite this fact to rebut *tomorrow's* skeptical questions. But I shall clearly be no better placed tomorrow than I am *today*, so there is no such fact.

And clearly enough, the skeptical argument doesn't essentially concern me, or the sign '+,' so it applies to all other language users and all other expressions. We have the skeptical paradox in full generality: "There can be no such thing as meaning anything by any word. Each new application we make is a leap in the dark; any present intention could be interpreted so as to accord with anything we may choose to do" (Kripke, 1982, p. 55).

Kripke's Wittgenstein commends a skeptical solution. The first part of the skeptical solution agrees with the skeptic that there are no facts described or misdescribed by meaning-ascriptions, but says: that doesn't matter, because such statements are *not aimed at stating facts*, but have a quite different, non-fact-stating role. Kripke seeks to make plausible his attribution of this idea to Wittgenstein by linking it to Wittgenstein's abandonment of the truth-conditional theory of meaning found in his *Tractatus* in favor of the conception of meaning as use advocated in his later writings, according to which an account of the use of a declarative sentence will comprise, in Kripke's view, a description of the conditions in which it may be appropriately asserted, together with an explanation of its role in surrounding linguistic and non-linguistic practices. The second part of the skeptical solution brings in the community. The conditions in which it is appropriate to say things like 'Jones means addition by "+"' are essentially communal – the remark is appropriately made when we have found that Jones makes statements using '+' which are in good agreement with the things we are ourselves inclined to say. The point and role of such remarks is to acknowledge him as a fully paid-up member of the community of adders, to convey that he can be relied upon not to come up with bizarre answers (like '5') to addition problems (like '68 + 57 = ?'), and so on.

Kripke's Wittgenstein thinks community involvement is essential to provide for the normativity of meaning. If we just consider Jones on his own, all there is is his inclination to apply the word in a certain way (to respond, unhesitatingly but blindly, to addition questions with certain answers); there is nothing for his usage to be in or out of accord with. There is no room, at the level of the isolated individual's use, for the crucial distinction between what *seems* to him right and what *is* right; so that we cannot speak of right at all (cf. Wittgenstein, 1967, §258). It is only when we bring in the community, and with it the possibility of agreement and disagreement between his use and that of the rest of us, that there can be a question of his applying the word rightly in a particular case. It is essential to realize that Kripke's Wittgenstein is not proposing that there is, after all, a fact constituting my meaning + by '+,' but an essentially communal fact: if that were his position, it would clearly be vulnerable to a community-wide version of the skeptical argument, for there would then be no less a problem about what rule the community is following than there is about the individual – there is but a finite stock of previous uses of '+' by the community, and that no more determines what function was meant than does the individual's past usage, and so on.

3 Is Semantic Irrealism Incoherent?

Kripke's argument has, quite justly, received a great deal of critical attention, mostly aimed at making out that the skeptical paradox admits of a straight solution – either one that Kripke overlooks altogether, or one that he considers but fails to rule out. These attempts may be divided into two broad groups. In the first come those which accept the assumption to which, notwithstanding his early insistence that there are to be "no limitations … on the facts that may be cited to answer the sceptic" (Kripke, 1982, p. 14), Kripke himself appears

to subscribe, that putatively meaning-constitutive facts must be specifiable in non-semantic, non-intentional terms. The main contenders here – aimed at a naturalistic solution – have been attempts to uphold some more or less sophisticated version of dispositional theory, or to show that a broadly causal account of meaning and/or reference escapes the skeptical argument. It has also been claimed that even if Kripke's objections are effective against a dispositional account, they do not dispose of the view that an expression's having a certain meaning consists in its being associated with an appropriate capacity.[3] Others – the second group – take issue with what they see as a substantial reductionist assumption underpinning the skeptical argument, and have accordingly sought to defend the view that semantic facts, or closely related facts about intentions, need not be reducible to facts of some other, naturalistic kind.[4]

Kripke himself describes the skeptical conclusion as "insane and intolerable" (Kripke, 1982, p. 60). But he believes that it is nonetheless a conclusion we have to accept. The skeptical solution, he hopes, enables us to do so. Others have taken a less optimistic view of the skeptical solution, arguing that the skeptical conclusion not only appears to be but really is intolerable. If they are right then there must be something wrong with the argument to it: a straight solution of some sort must be possible. Space does not permit detailed evaluation of the various alternatives which have been canvassed here. In this section I shall, instead, examine some arguments designed to establish the incoherence of the position to which Kripke is led by the skeptical argument. First, however, it will be useful to make some remarks about what that position – semantic irrealism, as I shall call it – involves.

Meaning-statements are made by means of declarative sentences. As such, they may be asserted on their own, and they may equally figure as components in conditionals, disjunctions, and other compounds. This is, arguably, by itself enough to ensure that they may with equal propriety be embedded in such contexts as 'It is true that …' and 'That … is a fact.' It might perhaps be insisted that such embeddings are acceptable only if the embedded sentences express claims which are subject to standards of correctness. I am not myself convinced that that is so, but even if it is, a proponent of the skeptical solution could hardly object on that score to our saying, for example, 'That Jones means addition by "plus" is a fact.' The aim of the skeptical solution is to rehabilitate (talk of) meaning in the face of the skeptical conclusion, by explaining how we can properly and correctly assert things like 'Jones means addition by "plus."' There is no unavoidable error in ordinary talk of this kind – Kripke is not advocating an 'error theory' of meaning discourse, analogous to John Mackie's error theory of ethical discourse (see Chapter 20, REALISM AND ITS OPPOSITIONS, §4): the error lies, rather, in prevalent philosophical (mis)interpretations, which construe meaning-statements as genuinely fact-stating or descriptive, having genuine truth-conditions.

It is, then, an obvious thought that for this very reason, it cannot be right simply to deny without qualification that meaning-statements are ever true, or that they state facts. Kripke anticipates such an objection to the skeptical solution's endorsement of the skeptical conclusion, and suggests that it may be defused by appeal to the 'redundancy' theory of truth (Kripke, 1982, p. 86. See Chapter 21, THEORIES OF TRUTH, §8). His thought seems to be that, since 'it is true that S means that p' has the same content as 'S means that p,' we are doing no more in asserting the former than we are in asserting the latter, and so are saying nothing from which a proponent of the skeptical solution need dissent. But this is puzzling.[5] In fact, the point seems to tell in precisely the opposite direction: just because, given a redundancy or deflationary conception of truth (and facts), 'It is true (is a fact) that S means that p' says no more than 'S means that p,' there can be nothing wrong with the former – and

accordingly, if the skeptical denial that meaning-statements are true or state facts is understood as involving this minimal notion of truth or fact, it must be wrong. The moral – apparently not clearly appreciated by Kripke – is that the skeptical conclusion, if it is to have even a chance of being acceptable, must be understood as invoking some more substantial conception of truth and facts. And if the skeptical solution is to have point, clarification of the more substantial notion(s) of truth and fact whose application to meaning-statements is to be denied becomes a matter of some urgency.[6] Whether a telling objection to Kripke can be erected around this point will be considered later. Meanwhile, I shall reserve the term 'true' for whatever more substantial notion might be taken to be in play, and employ 'correct' for the minimal sense. The skeptical conclusion can then be understood as claiming that meaning-statements are *never true*, but are (at best) *correct*.

More than one thinker has remarked upon the close similarity between the skeptical solution's combination of meaning irrealism with an attempt to rehabilitate meaning discourse by construing it non-descriptively, and more familiar projectivist attempts to save our thought and talk in other areas, such as morality, aesthetics, and modality, where the apparent absence of a suitable range of truth-conferring facts seems to preclude a fully realist construal of the discourse (cf. Wright, 1984b; Boghossian, 1989; 1990). And in one way, given the essentially normative character of the notion of meaning – on which all parties are agreed – together with the plausible claim that there can be no successful reduction of the normative to the purely factual, it may seem that meaning discourse is ripe for projectivist reconstruction (see Chapter 20, REALISM AND ITS OPPOSITIONS, §4). On the other hand, just because a projectivist treatment of some given region of discourse is a thesis about the kind of meaning attaching to statements belonging to it, it may be doubted whether a non-factualist or projectivist approach can coherently be applied to meaning itself. It may well seem that the philosophical point and advantage of, say, a projective treatment of ethical discourse (perhaps based upon the kind of expressivist reconstrual proposed by the emotive theory) would be substantially compromised, if coupled with the thesis that meaning discourse quite generally, and so any claim about the sort of meaning possessed by ethical statements in particular, is itself not genuinely factual but projective of, say, some attitude we have. These are, of course, no more than vague misgivings. Can they be transformed into a sharp and telling objection to the skeptical solution?

John McDowell wrote:

> It is natural to suppose that if one says 'There is no fact that could constitute its being the case that P,' one precludes oneself from affirming that P; … Given this supposition, the concession that Kripke says Wittgenstein makes to the sceptic becomes a *denial* that I understand the 'plus' sign to mean one thing rather than another. And now – generalizing the denial – we do seem to have fallen into an abyss: 'the incredible and self-defeating conclusion, that all language is meaningless' (Kripke, 1982, p. 71). It is quite obscure how we could hope to claw ourselves back by manipulating the notion of accredited membership in a linguistic community. (McDowell, 1984, p. 330)

The pessimistic conclusion is, however, too swiftly drawn. As we have seen, it is a condition of the coherence of Kripke's skeptic's argument that he is working with a substantial notion of fact, one for which the correctness of 'It is a fact that P' is precisely *not* guaranteed merely by the assertibility of 'P': that is, a more than merely deflationary or minimal notion of fact. But for this notion we can hardly expect there to be a generally unproblematic transition from 'It is not a fact that P' to denying that P. Kripke will want to hold, on the contrary, that there will be cases in which we can correctly assert that P, when it is not a fact that P. McDowell's

'natural supposition' just begs the question against him. It may be that irrealism about meaning, in contrast with irrealist theses in other areas, such as morals or mathematics, will turn out to suffer from some distinctive species of instability. But if so, further argument is needed to disclose it. Important arguments to the purpose have been advanced by Wright and Boghossian (cf. Wright, 1984a; Boghossian, 1989; 1990).

Wright argues in two stages: (1) irrealism about meaning leads to global irrealism and (2) global irrealism is incoherent or otherwise directly objectionable. Wright's globalizing argument pivots on what he calls the meaning-truth platitude, that 'the truth value of a statement depends only upon its meaning and the state of the world in relevant respects.' In its original version, from which this formulation of the platitude is taken, it runs thus:

> If the truth value of S is determined by its meaning and the state of the world in relevant respects, then non-factuality in one of the determinants can be expected to induce non-factuality in the outcome. (A rough parallel: If among the determinants of whether it is worth while going to see a certain exhibition is how well presented the leading exhibits are, then, if questions of good presentation are not considered to be entirely factual, neither is the matter of whether it is worth while going to see the exhibition.) A projectivist view of meaning is thus, it appears, going to enjoin a projectivist view of what it is for a statement to be true. Whence, unless it is, mysteriously, possible for a projective statement to sustain a biconditional with a genuinely factual statement, the disquotational schema '|P| is true if and only if P' will churn out the result that *all* statements are projective. (Wright, 1984a, p. 769)

Against global projectivism Wright advances several related but distinguishable considerations. An obvious worry, hinted at previously, is that a projectivist treatment of any particular class of statements has point only in so far as it draws a significant contrast between members of that class and other statements which are to be viewed as genuinely fact-stating, or apt for substantial truth. Relatedly, whilst a perfectly good distinction may *turn out* to be empty on one side, it may be doubted whether that could be an *a priori* matter, as would be the case with the needed distinction between fact-stating and non-fact-stating discourse if the globalizing argument is sound. Third, supposing the distinction satisfactorily drawn, the projectivist will surely want to regard it as a *discovery* that statements in the target class are non-factual – in particular, shouldn't the statement of the conclusion of the skeptical argument be *itself* genuinely factual? Fourth: there will be no *truths* about the (Kripkean) assertion conditions of any sentences, with the result that the premises of Kripke's version of the argument against private language (relating to the communally oriented character of the assertion conditions of meaning-statements), and hence also its conclusion, will enjoy a merely projective character (Wright, 1984a, pp. 769–770).

Boghossian agrees with Wright that irrealism about meaning inflates into global irrealism (though he gives a somewhat different argument for this claim), but he is not persuaded that this is intrinsically objectionable; instead, he argues that meaning irrealism leads directly to self-contradiction, independently of its implicitly global character. The argument[7] starts from a generalization of the point made above: that since any significant, declarative sentence is apt for truth in a merely deflationary sense, a non-factualist thesis about any class of statements cannot be understood as denying that any of those statements are true in that sense, but must be taken to involve a richer, more substantial notion of truth. The non-factualist is, as he puts it, "committed to holding that the predicate 'true' stands for some sort of language-independent property, eligibility for which will not be certified purely by the fact

that a sentence is declarative and significant" (Boghossian, 1989, p. 526). He then claims that a judgment that some sentence is or is not (substantially) true cannot but be a genuinely factual judgment. That is: the judgment that S is true (and likewise the judgment that S is not true) must itself be true or false, as opposed to being merely correct or incorrect. But the meaning non-factualist's distinctive thesis is that judgments about what a sentence means are not factual. Since what truth-condition a sentence possesses is a function of its meaning, it follows that judgments about what truth-condition a sentence has are likewise not factual. And since a sentence's having a particular truth-value cannot be a factual matter if its having a certain truth-condition is not, it further follows that a judgment about a sentence's truth-value can never be factual. Thus the meaning non-factualist is committed to denying that it is ever true that S is true. His position is thus self-contradictory.

Does either of these arguments succeed? Obviously enough, the crucial claim in Boghossian's argument is that the judgment that a sentence is (substantially) true must itself be genuinely factual. But it is anything but obvious that the meaning non-factualist must agree. He is, as we have seen, committed to the intelligibility of a thick (that is, more than merely deflationary) notion of truth; though whether he is further committed to its having a non-empty extension must, at this stage, be regarded as an open question. But acceptance of that much seems perfectly consistent with retention of the thin, merely deflationary notion which we are calling correctness. And so long as both notions are available, why can't the non-factualist hold, apparently with perfect consistency, that metalinguistic attributions of truth, falsity, correctness, and incorrectness are all alike, at most correct and never true? Boghossian believes that the non-factualist has not merely to make room for a thick (or as he says 'robust') notion of truth, but that he must *choose* between that and a purely deflationary one:

> It is an assumption of the present paper that the concept of truth is *univocal*… We should not confuse the fact that it is now an open question whether truth is robust or deflationary for the claim that it can be both. There is no discernible plausibility in the suggestion that the concept of a correspondence between language and world and the concept of a language-bound operator of semantic ascent might both be versions of the same idea. (Boghossian, 1990, p. 165, n. 17)

Clearly so crucial an assumption stands very much in need of supporting argument; surprisingly, Boghossian provides none, unless you think that the last sentence quoted does more than merely reassert what needs to be established.

This objection coincides pretty well, I think, with one of several developed in more detail by Wright (cf. Wright, 1992a, p. 234; 1993, pp. 318–324), who insists, as I have done, that the non-factualist is free to wield notions both of truth and correctness. Somewhat ironically, this distinction appears at first to provide the non-factualist with a ready way to interrupt Wright's own attempted *reductio* at the first, globalizing stage. As Wright's formulation makes plain, the final step involves an application of the Disquotation Scheme for 'true.' More specifically, he appears to have envisaged substituting the right- for the left-hand side of the scheme, to get from:

> '"P" is true' is not true

to

> 'P' is not true

But it now appears that the non-factualist may block this step: when 'P' is true, '"P" is true' will be merely correct, so that the Disquotation Scheme – 'P' is true if and only if P – fails right-to-left.

In fact, this claim relies upon a questionable assumption about the evaluation of conditionals, that is, that a conditional will hold (be true, or at least correct) only if there is no descent in value (from true to correct, say) between its antecedent and consequent. As against this, it may plausibly be claimed that we should require only preservation of designated value (where true and correct are designated, the remaining values not). However, whilst this makes it at least doubtful that the non-factualist can block the globalizing argument by rejecting the Disquotation Scheme outright, it leaves him with the resources for an equally effective rejoinder – indeed, a more satisfactory one, because it allows him to retain the Disquotation Scheme. For if the scheme is secured by adoption of the proposal that what is required for a conditional to hold is not that the consequent is true if the antecedent is, but only that designated values shall be preserved, then instances of the biconditional scheme will not support substitution of their components in complex contexts such as that involved in the globalizing argument.[8]

There is another, more obvious ground for dissatisfaction with the globalizing argument, at least in Wright's version. For it seems clear that, at least as formulated, the argument given works at best for a sense of 'statement' in which statements can be taken as *both* bearers of meaning *and* bearers of truth-value. A proponent of semantic irrealism need not deny that there is, or could be, such a sense of 'statement,' provided that he is granted a different sense in which statements have truth-values, but cannot sensibly be said to have meanings. Concerning any statement in this sense, he can claim that whether or not it is true *is* a factual matter; or more precisely, that its being so is not threatened by the non-factuality of meaning. It is true enough that whether or not a particular *sentence* is suitable for making a particular statement in this sense depends upon the sentence's having a certain meaning – and that, he holds, is not a factual matter. But the truth-value of a statement, in his preferred sense, does not depend upon the meaning of anything. It does not depend upon the meaning of the *statement*, because statements are not the sort of thing to have meanings; and it does not depend upon the meaning of a certain *sentence* – what depends upon a sentence's meaning being, rather, what statement(s) that sentence can be used to make. Thus non-factuality at the level of meaning does not induce non-factuality at the level of truth-value of statements.[9]

This discussion has inevitably been somewhat inconclusive. If what I have argued is right, it has not been shown that irrealism about meaning leads directly to contradiction, independently of its putative tendency to inflate into global irrealism; and it is at least open to question that it does globalize. And even if it does globalize in the way Wright and Boghossian both believe, it remains to be seen whether that leads to its collapse. Wright's arguments are suggestive of instability here, but appear less than decisive.

4 Wright on the Rule-Following Considerations

4.1 The Contractual Model of Meaning and Investigation-Independence

While Wright is sharply opposed to the semantic-irrealist conclusion which Kripke extracts from the rule-following considerations, he advances (Wright, 1980; 1981; 1984b) an argument (for which he claims Wittgensteinian origins) whose conclusion has – rightly or

wrongly – been seen as carrying implications for the notion of objectivity which are scarcely less radical, and no more palatable, than Kripke's. The argument is directed not at calling in question the very existence of facts about meaning, but at undermining what Wright takes to be an important misconception of their character. According to the conception under attack – the contractual model of meaning[10] – an expression's having a certain settled meaning consists in its being associated with a definite pattern of application which, once established, extends 'of itself' to new cases quite without any further assistance from us. Learning what the expression means is a matter of 'cottoning on' to such a pattern; our subsequent employment of the expression then either conforms, or fails to conform, with requirements already laid down, as it were, in the contract to which we have become party. Wright's contention is that, for reasons implicit in Wittgenstein's discussions of rule-following, the contractual model is fundamentally flawed and must be replaced by a conception of meaning as shaped by our ongoing use.

That is the immediate conclusion of Wright's argument. If for no other reason, the argument which purports to establish it deserves the closest scrutiny simply because the contractual picture is one which we may find both appealing and entirely natural, and which may, indeed, seem inevitable when we seek to understand what is involved in the normativity of meaning, so that Wright's conclusion is at the very least unsettling. But Wright draws a further conclusion which may appear not merely unsettling, but plainly intolerable. This concerns the way or sense in which ordinary factual statements may be held to be objectively true or false. What, in very general terms, we intend when we take a statement to be objectively true or false, is that its truth-value is in some way independent of our, or anyone else's, opinion. But this somewhat vague idea can be cashed out in various more specific ways. We ought not to be surprised if it should prove that what more precise characterization of it is found acceptable depends upon where one's sympathies lie in the dispute between realists and anti-realists in the theory of meaning (see Chapter 20, REALISM AND ITS OPPOSITIONS, §§1 and 2). Wright focuses upon one particular conception of objectivity – *investigation-independence* – which we might expect anyone of a realist persuasion to endorse. A realist, in Dummett's sense, about a certain class of statements – that is, one who holds that statements in that class are such that their truth-conditions may be fulfilled, or not, without our being even in principle capable of recognizing as much – evidently regards those statements as objectively true or false. But if a capacity for evidence-transcendent truth is taken as the criterion, the resultant sense of objectivity is very strong indeed. By their very nature, no effectively decidable statements will qualify as having objective truth-values in this sense. And the same, it is natural to suppose, goes for very many other perfectly ordinary statements which, though not effectively decidable in any strict sense, would normally be viewed as capable of objective truth. Assuming that the realist wishes to regard statements of these latter kinds as capable of objective truth, what alternative criterion should he adopt? Wright's plausible suggestion (cf. Wright, 1981, p. 99) is that he will embrace a notion according to which "confronted with any decidable, objective issue, there is *already* an answer which, if we investigate the matter fully and correctly, we will arrive at." For such statements, that is, objectivity of truth-value consists in the possession of a determinate, *investigation-independent* truth-value. But investigation-independence, Wright argues, requires the contractual model of meaning:

> Investigation-independence requires a certain stability in our understanding of our concepts. To think, for example, of the shape of some particular unobserved object as determinate, irrespective

> of whether or not we ever inspect it, is to accept that there are facts about how we will, or would, assess its shape if we do, or did, so correctly, in accordance with the meaning of the expressions in our vocabulary of shapes; the putative investigation-independent fact about the object's shape is a fact about how we would describe it if on the relevant occasion we continued to use germane expressions in what we regard as the correct way … The idea of investigation-independence thus leads us to look upon grasp of the meaning of an expression as grasp of a general pattern of use, conformity to which requires certain determinate uses in so far unconsidered cases. The pattern is thus to be thought of as extending of itself to cases which we have yet to confront. (Wright, 1981, p. 100)

If this is correct, a successful argument against the contractual model will be equally destructive of the idea that statements are capable of objective truth-value in the sense captured by investigation-independence.

It is obvious that any statement which is evidence-transcendently true or false will be objectively true or false in the sense captured by investigation-independence, but that the converse does not hold. In that sense, the latter is a weaker notion of objectivity than the former. And this, coupled with the fact that the notion of evidence-transcendent truth plays no part in the characterization of the weaker notion, might suggest that someone who rejects realism in Dummett's sense could endorse the claim that there are investigation-independent truths. But if Wright's argument is sound, this is an illusion. For the argument, as we shall see, makes essential use of an anti-realist premise, so that if he is unable to find fault with it elsewhere, the anti-realist must reject the notion of investigation-independent truth.

4.2 The 1980/1981 Argument

Can an individual speaker S, in her use of an expression E, defensibly be regarded as attempting to conform to a pattern of application, the requirements of which are already in place? Wright's argument, in its earlier version, divides the question into two.

First, can we defensibly regard S as aiming at conformity to such a pattern *independently* of the possibility of assessment of her performance by others?[11] The difficulty here is to see how it can be justified to describe the situation in terms of S's *recognizing* what her supposed pattern requires her to say, in any particular case, as opposed to her merely being *disposed* to apply E (or not, as may be). The former description is justified only if there is a distinction to be drawn between S's going on as the pattern demands on the one hand, and on the other her merely *seeming* to do so. But S cannot make this distinction for herself, since it is bound to seem to her that her sincere and considered application of E conforms to the requirements of the pattern; and by hypothesis, the distinction is not to be made out on the basis of others' assessment of her performance.

Since the contractual picture cannot be sustained for this case, we move to the question of whether it can make a difference to the situation if we add in, as it were, facts about the agreement, or lack of it, between S and the rest of us over the application of E. Wright argues that it makes no essential difference. Here it is crucial to remember that the question at issue is not whether agreement with the community somehow provides the standard of correctness, but whether bringing in agreement, or lack of agreement, with the community affords a way of keeping the contractual picture in play. As Wright puts it, it is the question: "How does others' agreement with me turn my descriptive disposition into a matter of recognition of conformity with a pattern, recognition of an

antecedent fact about how the communal pattern extends to the new case?" (Wright, 1981, p. 103). The answer, unstated but clearly implied, is that it cannot do so.

Wright restates this last part of the argument in a somewhat different way, which is worth noticing because it corresponds rather more closely to his later, and much terser, formulation (in Wright, 1984b). If S's agreement with the rest of us somehow made the crucial difference, so that she could be thought of as recognizing what the shared pattern dictates in a given case, then it should at least make sense for her to claim, should she find herself at loggerheads with the rest of us over the application of E, to recognize that we have gone off track. But the only proper conclusion for S to draw, given that she can find no way to persuade us that we have broken faith with our antecedent pattern, is – or so Wright contends – that she does not (and perhaps never did) know what E means (as we employ it). But if no one can recognize that the community has *gone off* the rails, no one can recognize that the community has *stayed on* them; mere lack of disagreement with the community cannot substantiate the claim to recognize what its supposed pattern requires.

It is tempting, as Wright notes, to think that "a solicitable community of assent just does make the relevant difference" (Wright, 1981, p. 104). But he gives a supplementary argument (cf. Wright, 1980, pp. 219–220) which, if good, shows that the temptation must be resisted. On the contractual model, the bearing of communal agreement over the application of E on the correctness or otherwise of S's use has to be understood in a quite particular way. It is not that communal agreement is *constitutive* of correct use. Correctness must consist in conformity with the requirements of the community's pattern, and communal agreement can be at best[12] good inductive evidence for that. In other words, on the contractual model, a community of assent on what should be said in a given case provides the standard against which individual applications of E are to be assessed *only because and in so far as* communal agreement can be taken to be based upon *recognition* of what the community's shared pattern requires. But once this is seen, it should be clear that we have no progress: so, far from answering the objection previously urged against the picture of individual, community-independent conformity to an antecedent pattern, bringing in the community merely shifts the target. For what now requires justification is description of the situation in terms of the community's *recognizing* what its supposed pattern requires, in any particular case, rather than in terms of its merely being *disposed*, collectively and non-collusively, to apply E (or not, as may be). The former description is justified only if there is a distinction to be drawn between the community's going on as its pattern demands on the one hand, and on the other its merely *seeming* to it that it is doing so. As Wright puts it:

> If 'correctness' means ratification-independent conformity with an antecedent pattern, there is apparent absolutely nothing which we can do to make the contrast active between the *consensus description* and the *correct description*. (Wright, 1980, p. 219, my emphasis)

Of course – and as Wright agrees – we may as a community retrospectively judge that our erstwhile, communally agreed verdict on a particular case was mistaken; but this can give no comfort to the contractualist, since it is obviously wholly tendentious to view this as a matter of our belatedly recognizing that we previously broke faith with the requirements of an antecedently determinate pattern.[13] The necessary contrast between recognizing what our pattern required, and our earlier, collective disposition concerning what to say merely changing, is evidently no less problematic than that between recognizing what our pattern requires us to say now and our present disposition.

Although it will scarcely have escaped the notice of readers already familiar with this debate, it is worth underlining the argument's reliance, in its closing step at least, if not earlier, upon an anti-realist premise to the effect that there is no sense to the claim that we operate with a distinction – in this case between the supposedly ratification-independent requirements of our pattern of use and how we think we should apply the expression in question – if there is nothing we can do to manifest a grasp of it. Wright himself is under no illusions, of course, about the need for such a premise, and indeed, stresses the point:

> If those arguments [i.e., the general anti-realist arguments against the intelligibility of attributing grasp of concepts of which there is no distinctive manifestation] are rejected, then there is … no obstacle to embracing the investigation-independence of decidable statements. If, and only if, one admits the need to describe how an understanding could be *revealed* of what it is for our consensus verdict … to fit the alleged investigation-independent fact of the matter … will one feel pressured to reject the 'double-element' conception. (Wright, 1980, p. 221)

That concludes my summary of Wright's argument in its earlier formulation. It leaves us facing three main questions: (1) Is the argument sound? (2) Is the rejection of investigation-independence it enjoins tolerable? (3) If the contractual model is to be scrapped, what should replace it?

4.3 Horrified Reactions

In view of the apparent innocence of the notion of investigation-independence, it is no surprise that others have seen its rejection not as a salutary corollary of the argument against the contractual model, but as revealing that something must have gone badly wrong, either in the argument Wright builds upon Wittgenstein's discussions of rule-following or in the rule-following considerations themselves. Thus John McDowell writes:

> If Wittgenstein's conclusion, as Wright interprets it, is allowed to stand, the most striking casualty is a familiar notion of objectivity. The idea at risk is the idea of things being thus and so anyway, whether or not we choose to investigate the matter in question, and whatever the outcome of any such investigation. That idea requires the conception of how things could correctly be said to be anyway – whatever, if anything, we go on to say about the matter: and this notion of correctness can only be the notion of how the pattern of application that we grasp, when we come to understand the concept in question, extends, independently of the actual outcome of any investigation, to the relevant case. So if the notion of independent-investigation is to be discarded, then so is the idea that things are, at least sometimes, thus and so anyway, independently of our ratifying the judgement that that is how they are. It seems fair to describe this extremely radical consequence as a kind of idealism.[14]

Although McDowell thinks that we cannot accept Wright's conclusion, and is thus committed to denying the *soundness* of the argument leading to it, he does not dispute its *validity*. In fact, he is committed to its validity, because he takes it to form the core of an effective 'transcendental argument against anti-realism' which reduces to absurdity the anti-realist premise upon which, as we have noted, the argument relies. It would, of course, be wholly tendentious, in the present context, to rest such a *reductio* on the alleged absurdity of the denial of investigation-independence itself. The absurdity lies, rather – or so McDowell contends – in the picture of language to which Wright's argument commits him, on which

there is no room for normativity, and so no room for meaning, at all. Of course, Wright does not himself think that his argument leads to this absurd conclusion. But there is no escaping it, McDowell claims, once we accept with Wright that at the individual level there is no going right or wrong in our use of words save in the context provided by communal assessment of individual use, and that as far as the community as a whole is concerned there is no authority to which its collectively agreed use is answerable, and no distinction to be drawn between the 'consensus description' and the 'correct description,' so that we cannot say that it 'goes right or wrong,' only that it 'just goes.' For this entails a picture of language use on which, 'at the basic level,' human beings are merely 'vocalizing in certain ways in response to objects,' no doubt to the accompaniment of certain 'feelings of constraint, or convictions of the rightness of what they are saying,' but at which 'there is no question of shared commitments – of the behavior … being subject to the authority of anything outside themselves … How, then, can we be entitled to view the behaviour as involving, say, calling things "yellow," rather than a mere brute meaningless sounding off?' And once we are committed to this picture of the 'basic' level, stripped of normativity altogether, there is no hope of reinstating it via the notion that individuals are subject to communal correction. As McDowell puts it,

> The problem for Wright is to distinguish the position he attributes to Wittgenstein from one according to which the possibility of going out of step with our fellows gives us the *illusion* of being subject to norms, and consequently the *illusion* of entertaining and expressing meanings.[15]

This attempt to turn Wright's argument on its head is, it seems to me, a complete failure. McDowell plainly takes it that when Wright observes that there is no standard against which the whole community's practice may be assessed, he is advancing this as his own view. But Wright is doing no such thing; he is himself offering a *reductio* of the idea that correct use is a matter of conformity with a ratification-independent pattern. McDowell appears entirely to have overlooked the crucial point that the conclusion that there is no distinction between the consensus verdict and the correct verdict is drawn on that hypothesis – hence Wright's conditional: 'If "correctness" here means ratification-independent conformity with an antecedent pattern, there is apparent absolutely nothing we can do to make active the contrast between the consensus description and the correct description.' Wright's argument as I understand it is that if communal correctness were a matter of conformity with such a pattern then, unless whether or not the community goes right is to be a verification-transcendent matter, there would have to be a distinction between the community's recognizing what its pattern requires and its merely thinking that it does. Since no content can be assigned to *this* contrast, there can be no content either to the distinction between the consensus verdict and the correct verdict, *on this supposition about what correctness consists in*. Given the obvious unacceptability of the conclusion to which it leads, we should reject that supposition.

Somewhat differently, Michael Dummett, in effect,[16] agrees with Wright that "an unflinching application of Wittgenstein's ideas about rules" leads us to deny that there can be predeterminate, investigation-independent facts; since he finds this conclusion incredible, he concludes that 'the "rule-following considerations" embody a huge mistake.' Wittgenstein was

> right to observe that, for the most fundamental of the rules that we follow, there is nothing *by which* we judge something to be a correct application of them. It certainly does not follow

> from this that, if we never do make such a judgement in some particular instance, there is no specific thing that would have been a correct application: to draw that inference, you need a general internalist premiss, that there is nothing to truth beyond our acknowledgement of truth.[17]

But this premise, Dummett complains, is totally implausible; to appeal to it in this context is simply to beg the question.

Since Dummett is discussing Wittgenstein, and not Wright, it would be unjust to complain of a failure to engage the latter's argument. But it is pertinent to observe that Wright's argument is, on the face of it, an argument of precisely the kind whose possibility Dummett denies, since it manifestly does not appeal to any premise to the effect that there is no more to truth than its being acknowledged.

The immediate reason why Dummett finds the rejection of investigation-independence incredible is that it appears to involve denying, in the case of an elementary calculation, that there is in advance of its being carried out any determinately correct result. Another reason that might be given (to which Dummett attaches great importance, both in the paper from which I have quoted and in earlier writings) is that it appears impossible to account satisfactorily for the value or usefulness of deductive inference without appealing to a distinction between a statement's being true and its being actually verified or recognized as true, and hence, it may seem, without invoking the possibility of investigation-independent truth. Indeed, this may be seen as a special case of a quite general difficulty. For it may seem that if Wright's conclusion stands, nothing approaching justice can be done to the conception we all have of human enquiry in general as a process of *discovery*. These are matters for genuine concern, to which – so far as I have been able to see – nothing in Wright's earlier presentations of the argument speaks. Pending explanation of how it might be alleviated, we have a strong motive for hoping, if not for suspecting, that there is after all a flaw in his argument.

4.4 Wright's Strengthened Argument

In fact, there is a flaw in it. The first part of the argument, aimed at showing that there can be no substance to the idea that an individual speaker is aiming at conformity to a pattern of use *independently* of the possibility of assessment of her performance by others, seems to me compelling. We should also agree that the contractual conception requires the possibility of community-wide (and so of near-community-wide) departure from its pattern for a given expression. But Wright's next claim – that the only conclusion a lone dissenter could properly draw, on finding herself unable to bring the rest of us round, is that she no longer understands the crucial expression, and so cannot be a competent critic of the rest of the community's use – is far from clearly correct. On the contrary, it appears that there is plenty of room for a proponent of the contractual view to resist it. No doubt I should be disconcerted to find the rest of the community lined up against me. But it is far from self-evident that, were this to happen, it must be that I have gone astray, and cannot be that *they* have done so. We can surely envisage circumstances in which the opposite would be the case. It is, for instance, at least conceivable that everyone else has, perhaps as a result of exposure to some insidious form of radiation which I have escaped, suffered eye or brain damage which makes red things look yellow to them, or which has somehow scrambled whatever neural assemblies are associated with their capacity to use color terms. Of course, this supposition is far-fetched; but it appears at least to make sense, and that is all that is required.

Reformulating the argument once again, Wright claims:

> none of us, if he finds himself on his own about a new candidate for φ-ness, *and with no apparent way of bringing the rest of us around*, can sensibly claim to recognize that the community has here broken faith with its antecedent pattern of application for φ; the proper conclusion for him is rather that he has just discovered that he does not know what φ means. (Wright, 1980, p. 218)

The italicized words are evidently crucial, since Wright's claim would clearly be preposterous without them. Even with them, the claim that a charge of community-wide error simply makes no sense would be unwarranted if the key words meant merely that there does not appear to be any way in which the isolated individual can persuade the rest that they have gone wrong. Earlier, Wright speaks of the individual being *incorrigibly* out of line, suggesting that there is, without qualification, no way in which he can bring the community around. But now, it seems to me, we need to be a lot clearer about just what supposition it is that we are being invited to entertain, before we can say what follows. Why is it that he can't do that? Are we also to suppose that the situation is, as it were, symmetrical – so that there is, equally, no way in which the community can bring the individual round to its way of thinking? If so, then unless it is further being assumed that one or other party is the victim of some cognitive malfunction which, however, mysteriously resists exposure, we are in effect being asked to suppose that the individual goes one way and the community the other, without either being cognitively at fault: but then it seems that the supposition effectively begs the key question, of whether or not there is a fact of the matter to be recognized.

Wright agrees that his earlier argument needs reinforcement at this point, and seeks to provide it in the later, refurbished version. This proceeds in two stages, corresponding to a distinction Wright draws between basic statements and others. Very roughly,[18] the former are statements involving only demonstratives together with concepts whose mastery consists in the possession of some appropriate recognitional capacity – concepts for which "competent use standardly presupposes no more than normal sensory capacities and ostensive teaching," such as concepts of color, taste, or pitch. In the first stage, Wright deploys a strengthened version of the argument we have been considering, restricted in scope to basic statements; the second stage then generalizes the conclusion – if basic statements lack objectivity of meaning, so must the remainder. The first-stage argument proceeds, in essentials, as before. But now Wright adds a supplementary argument to close off the gap opened up by the apparent possibility that the lone dissenter is right, the rest of his community having indeed gone astray in their application of basic concepts, perhaps as a result of the deleterious effects of some environmental contaminant upon their capacity to apply them reliably, or for some similar reason. What, in more detail, would it be like, he asks, for there to be available reason to think that everyone (else) had gone astray in their application of basic concepts?

Wright's argument starts from the idea that, if this is a genuine possibility, we should be able to see how a sustainable case could be made for thinking it to have been realized. We may suppose that the lone dissenter can put up a case of this sort: he points to (a) evidence that the rest have been exposed to a certain environmental contaminant and (b) evidence that when others have been exposed to this contaminant in the past their basic judgments of the relevant sort have been distorted. Against such a case, a doubt of the following kind may be raised: the evidence (b) involves the claim that the affected

subjects' basic judgments were distorted, and the basis for this claim is that those judgments were found to be at odds with basic judgments made by others who were not affected. The case assumes that we are warranted in taking it that the judgments made by those who were not affected were indeed correct. But might not those very judgments themselves have been the product of widespread error? Unless and until adequate reason can be provided to discount this possibility, the case the lone dissenter has sought to make is worth nothing. Wright's counter-claim is, in effect, that the possibility could only rationally be discounted by appeal to something like this principle, as being analytic of the notion of basic statement:

> If there is widespread non-collusive agreement on the truth of a basic statement S and there is adequate reason to suppose that the parties to this widespread agreement understand the concepts involved in S, and are functioning normally in normal conditions for exercise of the appropriate recognitional capacities, and there is no further evidence germane to the case, then anyone apprised of all these facts has adequate grounds for regarding S as true. (Wright, 1984b, p. 288)

The snag is that the objectivist about meaning can hardly regard this principle as analytic; on her view, it can be at best a contingent truth, and that will not be enough to see off the challenge. In short, the attempt to sustain the contractual model by appealing to the possibility of widespread communal error opens the doors to skepticism.

I shall not here try to evaluate this argument; Wright has, in my view, made a powerful case which – so far as I know – has yet to receive an effective reply. And if he is right, the case against objectivity of meaning relies, in its final form, on no specifically anti-realist premise. I leave the reader to ponder it, and turn instead to my second main – and by now, pressing – question.

4.5 Investigation-Independence and Objectivity of Judgment

The *term* 'investigation-independent fact' is indeed strongly suggestive of a familiar enough and, for all that it calls for philosophical articulation, seemingly indispensable notion of objectivity; so much so that the suggestion that we should deny that there are any such facts may strike us as the philosophical equivalent of red-rag-waving. I shall try to explain why the bulls should stand their ground.

Preparatory to introducing the notion of *objectivity of judgment* – as distinct from that of objectivity of meaning – Wright says:

> Cognition is *relational*: it is a matter of arriving at true opinions in a manner *sensitive* to states of affairs whose obtaining is somehow independent of one's so arriving. Moreover, such a sensitivity must be conceived as essentially fallible. (Wright, 1984b, p. 281)

Obviously the crucial words here are 'somehow independent.' In what does the independence of cognized states of affairs consist? Part of what is involved, at least, is that in any particular case where a subject S comes to know (and so forms a true opinion) that p, it should be the case that the state of affairs in virtue of which it is true that p does not depend in any way at all on S's coming to believe that p; or, indeed, on S's or anyone else's coming to hold any opinion on the matters in question. That is, we conceive of the relevant state of

affairs as such that its obtaining is consistent with universal ignorance of its doing so. That is one component in the notion of objectivity of judgment, as Wright characterizes it. This is naturally expressed in counterfactual terms: whenever a subject S is properly described as coming to know that p, it would (still) have been the case that p, even if neither S nor anyone else had investigated the matter, or formed any opinion on it. It seems to follow that endorsement of objectivity of judgment for a type of statement entails accepting that there are relevant states of affairs which obtain or not, independently of investigation, in this sense at least: it is the case that p (or not) independently of whether anyone ever did or will carry out an investigation to determine whether or not p (and *a fortiori*, independently of the result of any such investigation, were one (to be) carried out).

Taking in the other component which Wright includes in the idea of objectivity being charted – that is, the essential fallibility of judgment – the objectivist about judgment is committed to there being true claims of this sort:

(1) It is the case that p and it would (still) have been the case that p, even if no one had carried out an investigation to determine whether or not p, and even if someone had carried out an investigation, but one that issued in the verdict that not-p.

How about the case where we are concerned with some decidable, but as yet uninvestigated matter? What kind of claim should the objectivist about judgment make then? Well, suppose the question is whether some large integer *k* is or is not prime. Then the objectivist can say this:

k is either prime or not. If *k* is prime, then even if no one ever investigates, it is prime, and were anyone to investigate but come up with a different answer, she would be mistaken; and if *k* is not prime, then again, even if no one ever investigates, it is not prime, and were anyone to investigate but come up, etc.

Generally, where decidable but as yet uninvestigated matters are in question, the objectivist about judgment may register the sense in which they concern objective states of affairs by asserting an appropriate statement of the form:

(2) Either p or not-p. If p, then even if no one ever investigates, p, and were anyone to investigate but come up with any other answer, she would be wrong; and if not-p, then again, even if no one ever investigates, not-p, and, etc.

The crucial point for present purposes is that counterfactuals of the kind embedded in these claims are precisely *not* counterfactuals about *what expressions it would be (have been) correct to apply in certain circumstances*. As such, they contrast sharply with the kind of counterfactual in terms of which investigation-independence is characterized by Wright – "the [investigation-independent] fact about the object's shape is a fact about how we would describe it if … we continued to use germane expressions in what we regard as the correct way" (Wright, 1980, p. 216).

An objectivist about judgment can assert claims of both kinds of the forms (1) and (2). This marks one clear sense in which he can regard certain statements as being true or false in virtue of states of affairs obtaining, or not, independently of investigation. If that is right,

then we should question McDowell's right to the following transition, integral to the argument by which he persuades himself that a "familiar and intuitive notion of objectivity" requires the contractual conception of meaning:

> The idea at risk is the idea of things being thus and so anyway, whether or not we choose to investigate the matter, and whatever the outcome of any such investigation. That idea requires the conception of how things could correctly be said to be anyway – whatever, if anything, we in fact go on to say about the matter. (McDowell, 1984, p. 325)

For endorsement of conditionals of types (1) and (2) seems quite enough to hit off the idea of objectivity (of things being thus and so anyway …). But those conditionals – at least on the face of it – say nothing about how things could correctly be said to be.

It may be replied that this is a mere artefact of formulation: surely the objectivist ought not to make any bones about accepting these reformulations:

(3) It is true to say that p and it would (still) have been true to say that p, even if no one had carried out an investigation to determine whether or not p, and even if someone had carried out an investigation, but one that issued in the verdict that not-p.

(4) Either it is true to say that p or it is true to say that not-p. If it is true to say that p, then even if no one ever investigates, it is true to say that p; and if it is true to say that not-p, then again, even if no one ever investigates, it is true to say that not-p.

Well, of course he should accept them, since their acceptability is guaranteed by the equivalence of 'p' with 'it is true to say that p.' But the effect of securing McDowell's first transition by appeal to the Equivalence Thesis is simply to put in question the next transition in his argument; that is, from the second sentence, just quoted, to:

> and this notion of correctness can only be the notion of how a pattern of application that we grasp … extends, independently of any investigation, to the relevant case.

This transition would be good if, but only if, we could pass from (3) to:

(5) It is true to say that p and it would (still) have been correct to assert 'p,' even if no one had carried out an investigation to determine whether or not p, and even if someone had carried out an investigation, but one that issued in the verdict that not-p.

or something to that effect. For it is only if some such transition to a (counterfactual) claim about what words it would have been correct to use is allowable that endorsement of objectivity in the sense of the premise can be made out to involve commitment to the idea of a pattern of application (of some words) extending independently of investigation. But it should be quite clear that the object-linguistic counterfactual simply does not entail its metalinguistic counterpart. In short, McDowell's argument is vitiated by a simple equivocation on 'how things could correctly be said to be.' The second step in his argument is good only if this says something about what words it would be correct to use; the first is good only if it does not.

If what I have said is right, there is after all a gap, discernible by one who rejects objectivity of meaning, between the truth of a statement and its actual verification. Contrary to first

appearances, denying that there are investigation-independent facts (in the sense in which Wright does deny this) does not involve denying that, when we correctly perform an elementary calculation, the correctness of our result is independent of our performance. In that sense we can agree with Dummett that there is, in advance of our carrying it out, a determinately correct result. And more generally, when we make a valid inference from true premises to the conclusion that p, the truth of our conclusion does not wait upon our coming to it; it can, *without presupposing the contractual model*, be acknowledged that it would still have been the case that p, even if we had not drawn the inference. In that sense, by making the inference we acquire knowledge of a fact of which we had previously been ignorant but which was already there to be known. There is, to be sure, more to be said before we can lay claim to a satisfying explanation of the usefulness of deductive inference.[19] But this much is, it seems to me, enough to dispel the appearance that no such account can be forthcoming if the contractual model of meaning is abandoned.

5 Concluding Remarks

I have concentrated here on two discussions, both of which enlist Wittgenstein's rule-following considerations in support of radical and highly revisionary conclusions about the objectivity of meaning – conclusions which may appear to entail, and have been taken to entail, consequences for the objectivity of truth and judgment which are no less radical and revisionary. My principal concern has been to argue that, however unpalatable these conclusions – Kripke's semantic irrealism and Wright's rejection of the contractual model and investigation-independence – may seem, we have as yet no compelling demonstration of their unacceptability, and thus have no advance right to think that the arguments leading to them must be unsound. I should like to conclude with some remarks about how we should view the situation.

First, whilst we have as yet no decisive ground for thinking semantic irrealism unstable, it is quite another question whether any argument Kripke gives, or might have given, compels its acceptance. The greater part of Kripke's argumentation – and certainly the most convincing part of it – is directed against attempts to explain, in naturalistically reductive terms, what it is to mean something by an expression. To the extent that it is effective, it secures its effect by taking undisputed features of the concept of meaning – generality of application and normativity – and showing that they elude explanation on the proposed naturalistic basis. Clearly no argument of this kind could undermine a view according to which semantic or more generally intentional phenomena are irreducible. To establish semantic irrealism requires no less than a demonstration that indispensable features of the concept of meaning cannot be jointly instantiated. The closest Kripke comes to providing one is in his dismissive discussion of the idea that meaning something is a *sui generis* 'unique introspectible state'; but this comes nowhere near to a demonstration that severally essential ingredients in the concept of meaning cannot coexist. Furthermore, the required features are, as Wright reminds us, apparently coherently coexemplified in our standard intuitive notion of intention. We have no *a priori* guarantee that that notion could not turn out to be incoherent; but no reason to think it so is yet in sight.

Matters stand otherwise, it seems to me, with Wright's conclusion. Here we are faced with a *prima facie* compelling argument for the bankruptcy of the contractual model to which, so far as I have been able to see, no effective counter has been provided. And if I am

right, horrified reactions to the ensuing rejection of investigation-independence can be seen to be misplaced, once that notion is properly separated from a more modest notion of objectivity which can perfectly well survive without contractual underpinning. What does then become pressing is the need for a satisfying, detailed account of how we may view meaning as – in Wright's own, somewhat opaque phrase – "shaped by features of our ongoing linguistic behavior." This is one direction in which we have a good way yet to travel, before we can reckon ourselves to have appreciated the full significance of Wittgenstein's remarks on rule-following.[20]

Notes

1 See, in addition to the works cited, Peacocke (1981), McGinn (1984), Budd (1984), Pears (1988, chs 16–18), Williams (1991), Luntley (1991), and Wright (1989, pp. 239–245).
Another important focus of controversy which must be left unexplored here is the exact relationship between Wittgenstein's discussion of rule-following and his arguments against the possibility of a private language. See Chapter 11, MEANING AND PRIVACY.

2 This calls for some qualification: whilst there is, so far as I know, no published attempt to show that perverse Kripkean hypotheses break down when we try to work them through in detail, at least one critic – Neil Tennant – attempts to make such a case in *Taming of the True* (1997).

3 For dispositional theories see, for example, Papineau (1987), Fodor (1987; 1991). For doubts about the efficacy of Kripke's objections, see Blackburn (1984) and Forbes (1983). Attempts to uphold a causal account of meaning or reference in the face of the skeptical argument may be found in Goldfarb (1985) and McGinn (1984, pp. 164–166), who also defends the capacity proposal (pp. 168–175). Chomsky (1986) argues that Kripke improperly restricts the search for meaning-constitutive facts – the claim that you formerly followed a certain rule is a *theoretical* claim, and as such answerable to future evidence, as well as evidence concerning, for example, your past linguistic behavior or conscious mental life. Boghossian (1989) provides a useful survey and assessment of attempts at a straight solution which accept Kripke's reductionist assumption. See Chapter 8, A GUIDE TO NATURALIZING SEMANTICS.

4 For criticism of Kripke's reductionist leanings and proposals for non-reductive solutions, see Boghossian (1989, pp. 540–549), McGinn (1984, pp. 150–164), and Wright (1984a, pp. 772–777).

5 As some commentators, for example, Blackburn (1984, p. 285), have remarked.

6 I cannot, for reasons of space, pursue the matter here – for illuminating discussion, see Wright (1992a) *passim*. See also Chapter 20, REALISM AND ITS OPPOSITIONS, final section.

7 Short version, Boghossian (1989); full dress, Boghossian (1990).

8 The difficulty has been stated with respect to Wright's version of the globalizing argument – but, as Wright himself points out, somewhat similar troubles afflict Boghossian's version: cf. Wright (1992a, pp. 218–220).

9 Blackburn (1989) makes a similar objection to Boghossian's version of the argument. Doubts about the sufficiency of this kind of response to the globalizing argument are developed in Wright (1992a, pp. 222–226).

10 In the later paper (1984b) Wright refers to this doctrine, according to which meanings are conceived of as "fully settled by over-and-done-with behavioral and intellectual episodes," as that of *objectivity of meaning*.

11 Wright conducts this stage of the argument in terms of the question whether sense can be given to the claim that an individual is being faithful to an idiolectic pattern of application. But as I understand it, what really matters, for this part of the argument, is not whether she is viewed as seeking to suit her performance to a pattern peculiar to herself rather than a shared pattern, but whether others are supposed to be able to evaluate her performance.

12 At best, as Wright observes (1980, p. 219), it is far from clear how the probability of communal agreement being in or out of line with the requirements of the ratification-independent pattern could be assessed.

13 Cf. Wright:

> Of course, it may happen that the community changes its mind; and when it does so, it does not revise the judgement that the former view enjoyed consensus. But that is a fact about our procedure; to call attention to it is to call attention to the circumstance that we make use of the notion that we can all be wrong, but it is not to call attention to anything which gives sense to the idea that the wrongness consists in departure from a ratification-independent pattern. (1980, pp. 219–220)

14 McDowell (1984, p. 325). Whether this is a fair assessment of the import of Wright's argument, and whether, in particular, McDowell is right to identify the familiar intuitive notion of objectivity to which he gives expression with that of investigation-independence, are questions to which we must shortly return.

15 The scattered quotations from McDowell are all taken from his (1984, p. 336).

16 In effect, because Dummett is not explicitly discussing Wright's argument, though it may be that he does in fact have that argument in mind.

17 The quotations are from Dummett (1994, pp. 63–64).

18 This is a very rough description. For a much more careful account, see Wright (1984b, pp. 276–283). A needed further refinement of the notion of basic statement is provided in Wright (1992b, pp. 40–42).

19 For further discussion, see Dummett (1973; 1991a, pp. 36–42, 305–306; 1991b, ch. 7).

20 Thanks to Jim Edwards and Crispin Wright for very helpful comments.

References

Baker, G., and P. M. S. Hacker. 1984a. "On misunderstanding Wittgenstein: Kripke's private language argument." *Synthese*, 58(3): 407–450.

Baker, G., and P. M. S. Hacker. 1984b. *Scepticism, Rules and Language*. Oxford: Blackwell.

Baker, G., and P. M. S. Hacker. 1985. *Wittgenstein: Rules, Grammar and Necessity*. Oxford: Blackwell.

Blackburn, S. 1984. "The individual strikes back." *Synthese*, 58(3): 281–301.

Blackburn, S. 1989. "Wittgenstein's irrealism." In *Wittgenstein – Towards a Re-evaluation*, edited by R. Haller and J. Brandl, pp. 13–26. Vienna: Hölder-Pichler-Tempsky.

Boghossian, P. A. 1989. "The rule-following considerations." *Mind*, 93(392): 507–549.

Boghossian, P. A. 1990. "The status of content." *Philosophical Review*, 99(2): 157–183.

Budd, M. 1984. "Wittgenstein on meaning, interpretation and rules." *Synthese*, 58(3): 303–323.

Carruthers, P. 1984. "Baker and Hacker's Wittgenstein." *Synthese*, 58(3): 451–479.

Chomsky, N. 1986. *Knowledge of Language*. New York: Prager.

Dummett, M. 1973. "The justification of deduction." In *Truth and Other Enigmas*. London: Duckworth, 1978.

Dummett, M. 1991a. *Frege: Philosophy of Mathematics*. London: Duckworth.

Dummett, M. 1991b. *The Logical Basis of Metaphysics*. London: Duckworth.

Dummett, M. 1994. "Wittgenstein on necessity: some reflections." In *Reading Putnam*, edited by P. Clark and B. Hale, pp. 49–65. Oxford: Blackwell.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1991. "A theory of content." In *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Forbes, G. 1983. "Scepticism and semantic knowledge." *Proceedings of the Aristotelian Society*, 84: 223–237.
Goldfarb, W. 1985. "Kripke on Wittgenstein on rules." *Journal of Philosophy*, 82(9): 471–488.
Holtzman, S. H., and C. M. Leich, eds. 1981. *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul.
Kripke, S. A. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.
Luntley, M. 1991. "The transcendental grounds of meaning and the place of silence." In Puhl, 1991, pp. 170–188.
McDowell, J. 1984. "Wittgenstein on following a rule." *Synthese*, 58(3): 325–363.
McGinn, C. 1984. *Wittgenstein on Meaning*. Oxford: Blackwell.
Papineau, D. 1987. *Reality and Representation*. Oxford: Blackwell.
Peacocke, C. 1981. "Rule-following: the nature of Wittgenstein's arguments." In Holtzman and Leich, 1981, pp. 75–95.
Pears, D. 1988. *The False Prison*, vol. 2. Oxford: Oxford Unversity Press.
Puhl, K, ed. 1991. *Meaning Scepticism*. Berlin: De Gruyter.
Tennant, N. 1997. *The Taming of the True*. Oxford: Clarendon Press.
Williams, M. 1991: "Blind obedience: rules, community and the individual." In Puhl, 1991, pp. 93–125.
Wittgenstein, L. 1967. *Philosophical Investigations*, 3rd edn. Oxford: Blackwell.
Wittgenstein, L. 1978. *Remarks on the Foundations of Mathematics*, 3rd edn. Oxford: Blackwell.
Wright, C. 1980. *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.
Wright, C. 1981. "Rule-following: objectivity and the theory of meaning." In Holtzman and Leich, 1981, pp. 99–117.
Wright, C. 1984a. "Kripke's account of the argument against private language." *Journal of Philosophy*, 81(12): 759–778.
Wright, C. 1984b. "Rule-following, meaning and constructivism." In *Meaning and Interpretation*, edited by C. Travis, pp. 271–297. Oxford: Blackwell, 1986.
Wright, C. 1989. "Wittgenstein's rule-following considerations and the central project of theoretical linguistics." In *Reflections on Chomsky*, edited by A. George, pp. 233–264. Oxford: Blackwell.
Wright, C. 1992a. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.
Wright, C. 1992b. "Scientific realism and observation statements." In *Science and Subjectivity*, edited by D. Bell, pp. 21–46. Berlin: Academic Verlag.
Wright, C. 1993. "Eliminative materialism: going concern or passing fancy?" *Mind & Language*, 8(2): 316–326.

Further Reading

Blackburn, S. 1984. *Spreading the Word*. Oxford: Clarendon Press.
Fogelin, R. 1976. *Wittgenstein*. London: Routledge and Kegan Paul.
Putnam, H. 1981. "Convention: a theme in philosophy." *New Literary History*, 13(1): 1–14. Reprinted in *Realism and Reason: Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press, 1983.
Wright, C. 1989. "Review of McGinn, *Wittgenstein on Meaning*." *Mind*, 97(390): 289–306.

Postscript: Factualism and New Problems for Rule-Following

DANIEL WEE

This postscript attempts to outline recent developments in the philosophical literature on rule-following. In recent years commentators have proposed radical interpretations of Kripke's Wittgenstein that take his position to be compatible with a factualist account of meaning ascriptions. Additionally, new problems concerning rule-following have emerged: Crispin Wright argues that there are serious problems for the idea that linguistic competence can be construed in terms of rule-following, and Paul Boghossian contends that certain intuitively plausible accounts of rule-rationalization and explanation are threatened by a vicious regress that leads one to adopt a problematic conception of blind rule-following.

1 Factualist Readings of Kripke's Wittgenstein

One of the more prominent developments in the literature on Wittgenstein's rule-following considerations has been the dispute between factualist and non-factualist readings of Kripke's Wittgenstein (henceforth KW). The traditional reading of Kripke's exposition of Wittgenstein's remarks on rule-following, as exhibited by Hale in the chapter above, construes KW as ultimately advocating a non-factualist account of ascriptions of meaning: the notion that sentences such as "Jones means *addition* by '+'" do not state facts or have truth-conditions. The reason often proffered for this reading is that KW is compelled to accept a skeptical conclusion that there are no facts that constitute an individual's meaning something by an expression, and then – as part of his "skeptical solution" – tries to avoid being committed to an error theory of meaning ascriptions by revoking the notion that meaning ascriptions are fact-stating.

However, this reading of KW has been challenged by an opposing view that takes KW's position to be compatible with the notion that ascriptions of meaning do state facts and possess truth-conditions. This factualist reading of KW has been proposed by a number of philosophers such as (Byrne, 1996; Davies, 1998; Kusch, 2006). Due to space constraints, we shall only be able to sketch a factualist reading proposed by George Wilson (1994; 1998; 2006).

Wilson's reading of Kripke (1982) envisions two interlocutors in a debate on the content and justification of meaning ascriptions. The first is a skeptic who puts forward a skeptical argument and the second is KW, who attempts to defuse the skeptic's radical conclusion via a skeptical solution. But while this rudimentary picture can be gleaned from the traditional readings of Kripke (1982), the matter of how exactly the skeptical solution is a solution to the skeptic's argument is less clear. In particular, commentators have struggled to pinpoint the skeptical claim that KW concedes to his skeptic and the aspect of the skeptical argument that KW tries to counter in his skeptical solution. One of the virtues of Wilson's reading is that, whether or not one accepts his factualist construal of KW, it nevertheless brings into sharp relief what KW and the skeptic do and do not have in common.

We shall start by laying out Wilson's construal of the skeptical argument. Given that Kripke's discussion focuses on functional expressions (e.g., "+" meaning the *addition* function) we will be characterizing the argument in relation to such expressions. However, it is worth remembering that KW's skeptic is advancing a skeptical argument against meaning ascriptions in general, and so analogous arguments can be deployed for other kinds of expressions such as sentences and predicates.

The first premise of the skeptical argument is a condition of meaning imposed by "Classical Realism":

> **Classical Realism (CR)**: If a speaker *S* means something by a functional expression "F," then there is a function *f* that has been adopted by *S* as the standard of correctness for her application of "F."

According to (CR), an individual means something by an expression only if she has associated her use of that expression with some extra-linguistic entity as the standard that governs the correct application of that expression. The extra-linguistic entity concerned differs for the type of expression: for functional expressions it will be a function, for predicates it will be a set of properties, and for sentences it will be a possible fact or state of affairs. As we shall see, this is the crucial premise that Wilson takes KW to be rejecting in his skeptical solution.

The second premise is the "grounding constraint":

> **Grounding Constraint (G)**: If there is a function *f* that has been adopted by *S* as the standard of correctness for her applications of "F," then there must be facts about *S* that fix *f* as the standard that she has adopted for her application of "F."

According to Wilson, the candidates for meaning-facts that KW's skeptic considers and finds fault with are facts that are supposed to satisfy the grounding constraint. The perceived failure to find such facts even under idealized epistemic conditions thus leads the skeptic to draw what Wilson calls the "basic skeptical conclusion":

> **Basic Skeptical Conclusion (BSC)**: There are no facts about *S* that fix any function *f* as the standard of correctness she has adopted for her application of "F."

Thus, the skeptic finally draws his "radical skeptical conclusion":

> **Radical Skeptical Conclusion (RSC)**: No one means anything by a functional expression "F."

According to Wilson, KW's skeptic is committed to the argument from (CR), (G), and (BSC) to the conclusion (RSC). The difference between KW and the skeptic, however, is that KW sees (RSC) as "the incredible and self-defeating conclusion, that all language is meaningless" (Kripke, 1982, p. 71), and so KW endorses the contrapositive argument from ¬ (RSC), (G), and (BSC) to the conclusion ¬ (CR). Given Wilson's structuring of the arguments, it is clear what is shared and disputed between KW and his skeptic: KW and his skeptic both accept (BSC), but KW rejects (RSC) along with the classical realist condition for meaning (CR) that leads to it.

The way is now open for Wilson to argue that KW's skeptical solution can be compatible with factualism about meaning ascriptions. Under the classical realist account, in order for a meaning ascription like "Jones means *addition* by '+'" to be true, there must be facts in virtue of which it is true, and these facts must be facts concerning Jones's relation to the *addition* function as the standard of correctness for her use of "+." But accepting that there are no such facts (as (BSC) claims) is only to accept that there are no meaning-facts under the classical realist construal of what meaning-facts are supposed to be. By rejecting (CR),

KW can still make room for the notion that meaning ascriptions are true in virtue of facts, just that these facts would not be classical realist facts.

Of course, this is only an outline of Wilson's argument for the claim that KW's skeptical solution is compatible with factualism about meaning ascriptions. More needs to be done to show that there are such non-classical realist facts in virtue of which meaning ascriptions can be true. Nevertheless, philosophers have taken Wilson's argument seriously on both its exegetical and its philosophically substantive merits, and the debate between factualist and non-factualist construals of KW is ongoing. For examples of critical replies to Wilson see Miller (2010), Kremer (2000), and Soames (1998).

2 Wright's and Boghossian's Problems from "Blind" Rule-Following

The problem of rule-following expressed in KW's rule-following considerations is that of accounting for someone's internalization or acceptance of a rule. This problem has traditionally been considered the central problem of rule-following, and a wealth of literature has spawned in response to it. However, in recent years Crispin Wright (2007; 2012) and Paul Boghossian (2008; 2012) have independently identified further problems for rule-following that they claim remain even if we have a solution to KW's problem of rule-acceptance. These problems relate to the notion of "blind" rule-following, where we act on certain rules without being able to provide rational justification for why our actions are in accord with them.

2.1 Wright's Challenge to Construing Linguistic Competence as Rule-Following

According to Wright, a further problem emerges for the notion that linguistic competence can be construed in terms of following rules. To see how this problem emerges, Wright contends that in all cases of rule-following, it must be possible to structure the acceptance of the rule and the responses to it in terms of a *modus ponens* model. We can illustrate this model as follows in the case where a chess player makes the judgment that she may now castle:

> **Castling rule**: If neither the King nor one of its Rooks has moved in the course of the game so far, and if the squares between them are unoccupied, and if neither the King nor any of those squares is in check to an opposing piece, then one may castle.
> **Observation**: In this game neither my King nor this Rook have yet been moved, the squares between them are unoccupied, and neither the King nor any of those squares is in check to an opposing piece.
> **Judgment**: I may castle now.

However, Wright thinks that in order to apply the *modus ponens* model to a purported case of rule-following, what constitutes the major and minor premises must be extricable from each other. By "extricable," Wright means that the acceptance of the rule (the major premise) and the observation that the rule's antecedent is satisfied (the minor premise) can figure as independent states of mind. This requirement is imposed by the *modus ponens* model itself,

in order to distinguish between the general rule and the particular circumstances in which it is applied, so that "what properly belongs to the rule corresponds to the conditional major premise […] while what corresponds to the situational input will be given by the minor premise" (Wright, 2007, p. 492).

In the case of the castling rule the extricability condition is easily satisfiable: an individual can describe the positions and interactions of the pieces on the chess board as mentioned in the antecedent of the castling rule without any understanding of the castling rule itself. For instance, there can be chess novices who learn the castling rule only after they have learnt other chess concepts, and so are still able to recognize that neither their King nor Rook has moved, and so on, without having any inkling of the castling rule.

However, Wright thinks that the extricability condition for the *modus ponens* model cannot be satisfied in cases where one attempts to follow rules that govern the use of primitive or basic linguistic expressions, such as those involving "the simple classification of colours, or tastes, or Lockean secondary qualities generally" (Wright, 2007, p. 490). Wright characterizes such cases as instances of "blind" rule-following, "where nothing takes place which can naturally be regarded as *working out* what a rule requires" (p. 490). To illustrate how such a case would fare, Wright considers the example of following the rule for applying "red":

> **Rule for applying "red"**: If … x …, it is correct to predicate "red" of x
> **Observation**: … x …
> **Judgment**: It is correct to apply "red" to x.

In this case, unlike the castling rule, the grasp of the rule for applying "red" is inextricable from the ability to recognize when the antecedent of that rule is satisfied, because the ability to recognize the satisfaction of the antecedent requires mastery of the concept of *red* and hence grasp of the rule for applying "red" itself:

> To conceive of predications of 'red' as rule-governed in the manner of the model accordingly requires an anterior concept, '… x …,' whose satisfaction determines an input as appropriate for the application of the rule. But now it stares us in the face that this concept can hardly be anything other than: *red*! (Wright, 2007, p. 495)

So we arrive at a dilemma: either we maintain that in cases of following rules for using primitive expressions we can extricate the observation of the satisfaction of a rule's antecedent from acceptance of that rule, or we give up the idea that they are extricable and so discard the notion that competent use of primitive linguistic expressions can be construed in terms of rule-following.

If we accept the first horn of the dilemma then Wright believes we are committed to thinking of linguistic competence in terms of the Augustinian picture of language, according to which an individual can possess mastery of a concept anterior to the grasp of the rule of using that concept's expression. Wright finds this horn of the dilemma the worse of the two, and notes that it is something with "which the *Philosophical Investigations* begins and from which, as I put it, that text is a 'journey of recoil'" (Wright, 2012, p. 382). Given this, he is compelled to propose that we cannot construe competence in using primitive expressions in terms of rule-following. And since competence in using primitive expressions forms the basis for competence in using more complex expressions, in the sense that the

latter must be analyzed in terms of the former, it thus appears that we cannot construe linguistic competence in general in terms of following rules.

2.2 Boghossian's Problem for Rule Rationalization and Accepting Rule-Blindness

According to Boghossian, even if we have neutralized KW's problem of rule-acceptance, a further problem looms where we encounter a vicious regress in any attempt to explain or rationalize an individual's action as an instance of rule-following. To set up Boghossian's problem as such, we can help ourselves *pro tem* to the idea that KW's problem has been solved and so we are entitled to hold a non-reductionist view of rule-acceptance that Boghossian calls the "Intention View." This is the view that accepting a rule consists in forming an intention with general (prescriptive or normative) content, such as the intention to perform a certain action under certain circumstances. Boghossian characterizes the "Intention View" as just a specific form of the more general "Intentional View" according to which rule-acceptance consists in some intentional state or other – not necessarily an intention (Boghossian, 2008, p. 485).

Now Boghossian shares with Wright the conviction that all genuine cases of rule-following can be modeled according to a structure much like *modus ponens*. Given this and the Intention View, we can illustrate the case of someone following the rule 'Answer any email (that calls for an answer) immediately on receipt!':

Intention: For all x, if x is an email and you have just received x, answer it immediately!
Premise: This is an email that I have just received.
Conclusion: Answer it immediately!

What seems obvious here is that rule-following always involves inference: to follow a rule is to infer from one's acceptance of a rule and belief that the antecedent conditions of the rule are met, to the conclusion that the judgment that the action called for by the rule is now permissible or required. However, a vicious regress starts to stir when we realize that in performing an inference we are also following a rule of inference – one that can be characterized as:

(MP*) From 'If C, do A' and C, conclude 'do A'!

Consequently, it appears that in attempting to follow the rule for answering emails we need to make an inference, but making that inference itself requires following an additional rule of inference, MP*, and following that rule itself requires making an inference, and so on *ad infinitum*.

The upshot of Boghossian's considerations is that, even if we help ourselves to the Intention View of rule-acceptance, we still face a problem in attempting to rationalize or explain any individual's actions in light of a rule. This is because there seems to be an infinite number of cognitive steps between her acceptance of that rule and her inferring that the action called for by the rule is permissible or required on a particular occasion. It thus seems that in order to avoid such infinite regress, we must reject the Intention View, along with the more general Intentional View, and so concede that the acceptance of at least some

rules "cannot consist in the formation of a propositional attitude in which the requirements of the rule are explicitly encoded" (Boghossian, 2008, p. 495).

The rejection of the Intentional View is already problematic given that it seemed to constitute a promising response to KW's rule-acceptance problem. However, its rejection leads to a further problem of accounting for how someone can be said to follow a rule despite having no propositional attitude stipulating what the requirements of the rule are, and thus being unable to provide any rational justification for why her actions are in accord with the rule. This is a picture of rule-following as "blind" according to Boghossian, and it contrasts with a view where "rule-following is always fully *sighted*, always fully informed by some recognition of the requirements of the rule being followed" (Boghossian, 2008, p. 495). The challenge is thus to explain how rule-following can be "blind" in this highly problematic sense. For further discussion of Wright's and Boghossian's problems see Miller (2015).

References

Boghossian, P. 2008. "Epistemic rules." *Journal of Philosophy*, 105(9): 472–500.

Boghossian, P. 2012. "Blind rule-following." In *Mind, Meaning and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by A. Coliva, pp. 27–48. Oxford: Oxford University Press. DOI:10.1093/acprof:oso/9780199278053.003.0002.

Byrne, A. 1996. "On misinterpreting Kripke's Wittgenstein." *Philosophy and Phenomenological Research*, 56(2): 339–343.

Davies, D. 1998. "How sceptical is Kripke's 'sceptical solution'?" *Philosophia*, 26(1): 119–140.

Kremer, M. 2000. "Wilson on Kripke's Wittgenstein." *Philosophy and Phenomenological Research*, 60(3): 571–584.

Kripke, S. A. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

Kusch, M. 2006. *A Sceptical Guide to Meaning and Rules: Defending Kripke's Wittgenstein*. Chesham, United Kingdom: Acumen.

Miller, A. 2010. "Kripke's Wittgenstein, factualism and meaning." In *The Later Wittgenstein on Language*, edited by D. Whiting, pp. 167–190. Basingstoke: Palgrave Macmillan.

Miller, A. 2015. "Blind rule-following and the 'antinomy of pure reason.'" *The Philosophical Quarterly* (advance online publication). DOI:10.1093/pq/pqv023.

Soames, S. 1998. "Facts, truth conditions, and the skeptical solution to the rule-following paradox." *Philosophical Perspectives*, 12: 313–348.

Wilson, G. 1994. "Kripke on Wittgenstein on normativity." *Midwest Studies in Philosophy*, 19(1): 366–390.

Wilson, G. 1998. "Semantic realism and Kripke's Wittgenstein." *Philosophy and Phenomenological Research*, 58(1): 99–122.

Wilson, G. 2006. "Rule-following, meaning and normativity." In *The Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. C. Smith, pp. 151–174. Oxford: Oxford University Press.

Wright, C. 2007. "Rule-following without reasons: Wittgenstein's quietism and the constitutive question." *Ratio*, 20(4): 481–502.

Wright, C. 2012. "Replies part 1: the rule-following considerations and the normativity of meaning." In *Mind, Meaning and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by A. Coliva, pp. 379–401. Oxford: Oxford University Press. DOI:10.1093/acprof:oso/9780199278053.003.0015.

25

The Normativity of Meaning

ANANDI HATTIANGADI

1 Introduction

This chapter investigates the view that meaning is normative. *Meaning* is understood here in a broad sense to include such semantic properties as sense, reference, truth-conditions, content, and the like. *Normativity* can either be viewed as a property of representations or as a feature of the world. Normative representations include the judgment that you ought to give what you can to charity, the concept OUGHT[1] as it figures in that judgment, the statement 'all and only pleasure is good,' and the expression 'good' as it is used in that sentence. Sometimes, it is also held that there are normative facts and properties that correspond to normative representations, such as the fact that you ought to give what you can to charity or the property of goodness. The claim that meaning is normative can involve appeal either to normative representations or to normative facts. Indeed, it is possible to distinguish four versions of the claim that meaning is normative, two of each kind.

(1) *Meaning essentially involves normative judgment*: meaning something by an expression essentially involves following a rule, or making a normative judgment of some kind. (Cf. Brandom, 1994; Kripke, 1982; Ginsborg, 2011; 2012; Glock, 1996; Ichikawa and Jarvis, 2013; McDowell, 1984; Verheggen, 2011.)

(2) *Meaning is a source of normativity*: semantic truths imply or at least metaphysically necessitate normative truths.[2,3] (Boghossian, 1989; Ginsborg, 2011; Glock, 1996; Verheggen, 2011.)

(3) *Meaning facts are determined by normative facts*: the semantic facts supervene on the normative facts (perhaps together with the non-semantic, natural facts). (Cf. Brandom, 1994; Engel, 2000.)

(4) *Semantic concepts are normative*: the *concept* MEANING is normative, as are ascriptions of meaning. (Cf. Brandom, 1994; Gibbard, 2012.)

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Contemporary discussion of the normativity of meaning can be traced to Kripke's influential discussion of Wittgenstein's rule-following considerations (Kripke, 1982). Recall that Kripke asks what makes it the case that he means ADDITION rather than QUADDITION by 'plus,' where x quus $y = x$ plus y if x, $y \leq 57$, and x quus $y = 5$, otherwise. He argues for the surprising thesis that there is no fact of the matter what anybody means by any word. The idea that meaning is normative appears to play a crucial role in Kripke's argument (Boghossian, 1989; Hattiangadi, 2007; Soames 1997). For instance, Kripke rejects the view that meaning is determined by speaker dispositions on the grounds that "the relation of meaning and intention to future action is *normative*, not *descriptive*" (Kripke, 1982, p. 37). The problem, Kripke says, is "that my present mental state does not appear to determine what I *ought* to do in the future" (Kripke, 1982, p. 56). Since then, a number of philosophers have defended the view that meaning is normative in some sense or other (e.g., Boghossian, 1989; 2005; Bilgrami, 1992; Brandom, 1994; 2000; Gampel, 1997; Gibbard, 1994; 2012; Ginsborg, 2011; 2012; Glock, 1996; Ichikawa and Jarvis, 2013; Kusch, 2006; Lance and O'Leary-Hawthorne, 1998; Millar, 2002; Verheggen, 2011; Whiting, 2007; Zalabardo, 2012).

As Kripke's discussion makes abundantly clear, interest in the normativity of meaning is fuelled in large part by interest in foundational meta-semantics, what might be called the 'hard problem of intentionality.'[4] Are semantic properties identical to natural properties? Do the semantic facts supervene on the non-semantic facts? Can the semantic truths be reductively explained in non-semantic terms?[5] Kripke clearly suggests that the normativity of meaning has a crucial bearing on these questions. Indeed, he suggests that if meaning is normative, then meaning cannot be reductively explained, and concludes that there are no semantic facts. Though few contemporary philosophers adopt Kripke's skeptical conclusion (an exception might be Gibbard, 2012), many have similarly suggested that the normativity of meaning has a bearing on the hard problem of intentionality. The thought is that if meaning is normative, then there is a useful analogy to be drawn between the hard problem of intentionality and the hard problem of normativity – the problem of finding a place in the natural world for normative properties, such as rightness and goodness.

All four versions of the view that meaning is normative mentioned above have been presented as responses to the hard problem of intentionality. In the following, I will ask both whether meaning is normative in any of the four senses given above and if so, what bearing this has on the hard problem of intentionality. As we shall see, the idea that meaning is normative has very little bearing on the hard problem.

Before we proceed, it will be useful to highlight an important distinction that will be relevant to the discussion throughout. This is the distinction between norm-relativity and normativity proper. Norm-relative judgments concern whether or not something satisfies a norm or a standard.[6] Just about anything can figure as a norm or a standard: the school rules, the law, the requirements of rationality, the code of thieves, or a letter to Santa. To judge that wearing a uniform satisfies the school rules, or that squealing is forbidden by the code of thieves, is not to make a normative judgment at all. One way to distinguish normative and norm-relative judgments is in their rational relations to action. It is plausible that you are rationally required to intend to do what you judge that you ought to do in the genuinely normative sense. Yet you are not rationally required to intend to do what you judge to satisfy a norm or a standard in the merely norm-relational sense. A rational agent can judge that squealing is forbidden by the code of thieves, and yet fully intend to squeal.

Norm-relative truths can similarly be distinguished from genuinely normative truths.[7] On the face of it, the fact that wearing a uniform satisfies the school rules is not a normative

fact in any interesting sense; nor is the fact that squealing is forbidden by the code of thieves. On the face of it, the sentences we have used to state these truths merely describe the school rules and the code of thieves. These descriptions do not entail normative truths. Even if squealing is forbidden by the code of thieves, it does not follow that Bill the thief ought to avoid informing the police about his accomplices, or even that he has a *pro tanto* reason to do so. Of course, the fact that Bill will be punished if his accomplices learn that he has squealed may give him a *pro tanto* reason to avoid squealing. Nevertheless, the mere fact that squealing is forbidden by the code of thieves does not itself give him a *pro tanto* reason not to squeal.

The distinction between norm-relativity and normativity proper is crucial to the present discussion because it is only genuine normativity that generates the peculiar puzzles of metaethics. It is only if meaning is genuinely normative that there might be a fruitful analogy to be drawn between the hard problem of intentionality and the hard problem of normativity.

2 Meaning and Normative Judgment

Why think that meaning essentially involves normative judgment? One reason has to do with the familiar thought that language is a conventional activity, much like other conventional activities such as games (Wittgenstein, 1953): just as there are rules of chess, so too are there rules of language. Kripke picks up on this thought when he says that "[o]rdinarily, I suppose that, in computing '68 + 57' as I do, I do not simply make an unjustified leap in the dark. I follow directions I previously gave myself that uniquely determine that in this new instance I should say '125'" (1982, p. 10), and that such directions "must somehow be 'contained' in any candidate for the fact as to what I meant" (1982, p. 11).

The hypothesis that speaking a language essentially involves rule-following is a special instance of the more general thesis that meaning something by an expression essentially involves making a normative judgment of some kind. This is because, as is widely agreed, following a rule involves more than merely acting in accordance with it; it involves in some sense being guided by the rule, or at least accepting or endorsing the rule (Glüer and Wikforss, 2010). This in turn involves making a normative judgment of some kind.

Another motivation for thinking that meaning involves normative judgment derives from the thought that meaning is use. One version of the use theory of meaning, conceptual role semantics, claims that meaning something by an expression essentially involves the disposition to make certain inferences. For example, what it is to mean addition by 'plus' on this view is to be disposed to answer with the sum of any two numbers when asked (e.g., Block, 1986; Field, 1977; Harman, 1999; Peacocke, 1992). However, as Kripke has argued, the fact that you have this disposition is compatible with your meaning something deviant such as quaddition by 'plus.' In the face of this worry, normativists suggest that what it is to mean something by an expression is to be disposed to accept that certain inferential transitions are correct or appropriate. For instance, to mean addition by 'plus' on this view is to be disposed to judge that '125' is the correct answer to 'what is 57 plus 68?' (cf. Glock, 2005; Brandom, 1994).

Does meaning essentially involve normative judgment? There are some excellent reasons to think not. First of all, rules that might be thought to be constitutive of meaning are not the sorts of rules that can be followed (Glüer and Pagin, 1999; Glüer, 1999). Constitutive rules of meaning specify what a sentence means, and this does not naturally coincide with

telling speakers what to do. Consider, for instance, a rule that says: 'plus' means addition in English. This is a *descriptive* rule, which captures a generalization about usage. Since it does not tell anyone what to do, it is not at all obvious what is involved in following it. Indeed, even if you accept such a rule, your acceptance of it can only play the role of a belief in practical reasoning (Glüer and Pagin, 1999; Glüer and Wikforss, 2010). Furthermore, the judgment that some use of an expression accords with its meaning, or that it satisfies a constitutive rule of use is a norm-relative judgment, rather than a normative one.

Some have sought to sidestep this problem by formulating constitutive rules in terms of correctness conditions, and then arguing that correctness is a normative notion (Boghossian, 1989; 2003; Gibbard, 2003). The idea is that in order to mean addition by 'plus' it is necessary to accept or endorse a rule such as the following:

> *Plus-Rule*: $\ulcorner \varphi + \psi = \omega \urcorner$ is correct iff the number denoted by ω is the sum of the numbers denoted by φ and ψ.

If 'correct' is a normative term, having to do with what one ought to do or may do, then to accept *Plus-Rule* is to make a normative judgment of some kind (Boghossian, 2003; Gibbard, 2003). However, 'correct' is not a normative term across all contexts (Bykvist and Hattiangadi, 2013), and it is not a paradigmatic normative term like unqualified uses of 'ought,' 'may,' or 'reason.' So, to make sense of the idea that judgments of correctness have to do with what one ought to do or may do, we need a more precise characterization of what judgments about what one ought or may do are entailed by judgments of correctness.

What normative judgments are entailed by correctness judgments, then? Perhaps a normativist will say that in order to mean addition by 'plus' you must judge that you *ought* to respond with the sum of any two numbers when asked. But this view is implausible. Suppose that you discover that if you do not answer '5' when asked to add 57 and 68 in some circumstance, an evil demon will ensure that thousands will die. Though you will sensibly judge in this situation that you ought not to respond with the sum, we can hardly conclude that you do not mean addition by 'plus.' Or perhaps the normativist will say that to follow *Plus-Rule* you must judge that you *may* respond with the sum of any two numbers when asked (e.g., Whiting, 2009). But this view is equally implausible. For, in the circumstance just described, you sensibly judge that you ought not to respond with the sum, which is incompatible with the judgment that you may do so. The normativist could say instead that meaning addition by 'plus' requires that you judge yourself to have a *pro tanto* reason to respond with the sum of any two numbers when asked, where a *pro tanto* reason is a reason that can be outweighed by other considerations. However, this too is implausible. You might well think that you do not even have a *pro tanto* reason to respond with the sum in the case of the evil demon, yet we cannot conclude from this that you do not mean addition by 'plus' (cf. Glüer, 1999; 2001; Hattiangadi, 2006; 2007; Wikforss, 2001). Perhaps the normativist should say that correctness is a *sui generis* normative notion, not reducible to obligation and permission (cf. McHugh, 2014). He might then say that you must at least judge that responding with the sum is correct in view of what 'plus' means, even if you do not take yourself to have a *pro tanto* reason to do so. But it is difficult to see in what sense correctness is a normative notion if it is possible to coherently judge that some act is correct while simultaneously judging that you lack even a *pro tanto* normative reason to perform that act.

It might be tempting to respond to these cases by insisting that meaning addition by 'plus' essentially involves judgments of *semantic* correctness.[8] That is, in order to mean

addition by 'plus' you must be disposed to judge that it is *semantically* correct to respond with the sum of any two numbers, when asked. The trouble is that to judge that some answer is *semantically* correct is to judge that the answer is correct in view of what it means, or in view of satisfying *Plus-Rule*. And this is a *norm-relative* judgment, not a genuinely *normative* one. You are not rationally required to give the answer '125' merely because you judge that this answer satisfies *Plus-Rule* any more than you are rationally required not to squeal merely because you judge that squealing violates the code of thieves.

Perhaps the normativist will be tempted to recoup, and say that it is sufficient for meaning addition by 'plus' that you judge that you ought to respond with the sum when asked in most cases, even if sometimes, such as in the evil demon case, you do not. But this undermines the thought that these normative judgments are constitutive of what you mean. Let's say that *schmaddition* denotes a function that deviates from the addition function only in those cases in which you are morally required not to respond with the sum. In particular, let's say that in the evil demon case, the schmaddition function delivers the answer '5.' Clearly, the pattern of normative judgments you are typically disposed to make is consistent with your meaning schmaddition rather than addition by 'plus.' Indeed, more of the normative judgments you are disposed to make are consistent with your meaning schmaddition by 'plus' than are consistent with your meaning addition by 'plus.' Assuming that you do not mean schmaddition by 'plus,' what you mean cannot be determined by which normative judgments you are typically disposed to make.

Hannah Ginsborg (2011) puts a kind of particularist twist on the idea that meaning involves normative judgment. According to her, meaning something by an expression need not involve explicit acceptance or endorsement of any rule, only a kind of implicit rule-following. What it is to implicitly follow a rule for Ginsborg is to be disposed to make the appropriate pattern of normative judgments. For instance, on her view, it is necessary for meaning addition by 'plus' that a subject be disposed to judge that '57 + 68 = 125' is correct, that '125' is the answer one ought to give to '68 + 57 = ?' She describes this as a 'primitive' attitude in the sense that it need not be the result of conscious rule-following, or even the product of a *post hoc* realization of which rule one has been following (Ginsborg, 2011). Yet this particularist twist does not help to address the foregoing objections. Since judgments of correctness are not normative in every context, we need to know what normative judgments correspond to these primitive judgments of correctness. Perhaps Ginsborg will say that to mean addition by 'plus' one must be disposed to judge that one ought to or has a reason to answer with the sum, whenever one is asked to add any two numbers. But in the evil demon case, you are not disposed to judge that you ought to or even have a reason to answer with the sum, yet it does not follow that you do not mean addition by 'plus.' If Ginsborg attempts to relax the condition, and say that meaning addition by 'plus' involves making the relevant pattern of normative judgments most of the time, the pattern of normative judgments you typically make will not determine what you mean, since your dispositions will be compatible with the hypothesis that you mean schmaddition, or some other such bizarre function by 'plus.'

Claudine Verheggen (2011) claims instead that understanding the meaning of an expression or a sentence essentially involves the acceptance of certain instrumental oughts. On her view, it is essential to meaning addition by 'plus' that you accept that *if* you want to tell the truth, then you ought to say that 57 + 68 = 125 when asked. This suggestion avoids the above difficulties, because one can coherently accept this instrumental ought statement in the evil demon case, while accepting that all things considered, one ought not to give the

answer 125. The trouble is that instrumental ought judgments are merely norm-relative, not normative. 'If you want to get rich, you ought to be ruthless,' is norm-relative since it says that being ruthless is a necessary condition for satisfying the desire to be rich (against the background of the common knowledge of the context). Instrumental ought statements are not genuinely normative because they *describe* a necessary condition for satisfying a desire in context, and generally speaking, we do not think that statements describing necessary conditions are genuinely normative.[9] Thus, even if Verheggen is right that meaning essentially involves instrumental ought judgments, it does not follow that meaning essentially involves normative judgments.[10]

So far, we have considered internalist versions of the view that meaning involves normative judgments. Externalists deny that a speaker's private normative judgments are essential to what she means. Those externalists who hold that normative judgment is essential to meaning appeal instead to the normative judgments of the wider community, or communal practices in which those judgments are implicit. Such an externalist will be impatient to respond to the foregoing objections by appeal to a rule-governed practice, or some other such social mechanisms. For instance, Brandom argues that meaning something by an expression is constituted by inferential rules which are maintained by a practice of 'deontic scorekeeping' in which participants of the practice keep track of which inferences you are committed to or are entitled to make. For instance, in our practice, from a statement of 'x is red' you are committed to saying that 'x is colored' if asked; similarly, you will be permitted by our practice to say '57 + 68 = 125' and not permitted to say '57 + 68 = 5.' Now, you might on some occasion make a mistake, and forget to carry, but what you mean by 'plus' is not affected by this, since it is determined by the community-wide pattern of normative judgments. When you make such a mistake, it follows that you have said something that is not permitted by the semantic practices of your community. However, this sort of view fares no better with respect to the difficulties we have just canvassed than the views rejected previously. Just as you will sensibly judge in evil demon cases that you have no reason to respond with the sum, so too will many of your peers. Even if the whole community judges that you have no reason to answer with the sum in the evil demon case, it does not follow that you do not mean addition by 'plus' (cf. Hattiangadi, 2003).

It is time to move on to the question whether the hypothesis that meaning involves normative judgment has any bearing on the hard problem of intentionality. Recall that Kripke argues that dispositionalism fails because meaning is normative. He suggests the following argument: if meaning is normative, and normativity cannot be reductively explained, then meaning cannot be reductively explained either (cf. Boghossian, 1989; Gibbard, 2012; Kripke, 1982; Verheggen, 2011). But this is a *non sequitur*. The claim that meaning essentially involves normative *judgments* does not entail that there are any normative *facts*, yet the claim that normativity cannot be reductively explained concerns normative *properties* or *facts*. If meaning is constituted by normative *judgments*, and normative *facts* or *properties* cannot be reduced, it does not follow that meaning cannot be reduced. All that follows from the claim that meaning essentially involves normative judgment is that any reductive explanation of meaning must reductively explain normative judgment. But there is no *prima facie* reason to think that normative judgments are more difficult to reductively explain than non-normative judgments. Indeed, if there is a difficulty for a naturalist dispositionalist to reductively explain normative judgment, this difficulty seems to arise because normative judgments have *semantic properties*, not because the judgments are normative. That is,

if a dispositionalist account of normative judgment is unsatisfactory, this is because normative judgments have semantic properties, not because meaning is normative.

In contrast to Kripke, others appeal to the normativity of meaning in order to provide a positive, reductive account of meaning (e.g., Ginsborg, 2011; Brandom, 1994). These normativists grant that meaning cannot be reductively explained in natural scientific terms, but suggest that they can give an account of meaning that answers Kripke's problem in natural and normative terms. However, the claim that meaning essentially involves a normative judgment does not aid this reductive explanatory project either. The reason is that normative judgments themselves have semantic properties, so a reductive explanation that appeals to normative judgments in the *explanans* will be viciously circular. Recall that a reductive explanation of meaning is an explanation of meaning in *non-semantic* terms. Since normative judgments have semantic properties, they cannot figure in a reductive explanation of meaning.

Might the hypothesis that meaning involves normative judgment constitute a *partial* naturalistic solution to the hard problem of intentionality (e.g., Ginsborg, 2011)? Perhaps the reductive explanation of meaning proceeds in stages: first, we reductively explain meaning in terms of normative judgments, and then we give a reductive explanation of normative judgments. However, there is no obvious reason to think that the task of giving a reductive explanation of normative judgments is in any way easier than the task of giving a reductive explanation of linguistic meaning in the first place.

3 Meaning as a Source of Normativity

As we have seen, there is good reason to reject the view that meaning essentially involves normative judgment. However, there is a potentially more plausible idea in the vicinity. This is the idea is that (a) meaning is constituted by rules and (b) the rules constitutive of meaning are *sources of normativity*, that is, that they entail genuinely normative truths. This idea also goes back to the intuitive view that speaking a language is a conventional activity, much like playing chess. However, the thought is not that if you mean addition by 'plus,' then you *judge* that you ought to follow *Plus-Rule*. Rather, the thought is that if you mean addition by 'plus,' you *ought* to or have a *normative reason* to do what *Plus-Rule* requires of you.

Let us assume for the sake of argument that meaning is constituted by rules or requirements of some kind. This does not prejudge the question whether meaning is normative. We need only assume that constitutive requirements are 'property-requirements' in John Broome's sense (Broome, 2013). For instance, the rule in chess that states that the bishop ought to be moved only along the diagonal constitutes what it is to be a bishop in chess; it constitutes the property of being a bishop. In assuming this, we do not assume that the rules of chess are normative. By analogy, we can assume that *Plus-Rule* constitutes the property of meaning addition by 'plus' without assuming that *Plus-Rule* is normative. Similarly, there are requirements of the law, of rationality, of the code of thieves, of the school, and so forth. And with respect to any system of property requirements, we can ask the normative question. We can ask whether any such system of requirements is a source of normativity, that is, whether it is necessarily the case that one ought to do or has reason to do what the system of requirements requires of one.

Several proponents of the normativity of meaning have suggested that constitutive rules of meaning are sources of normativity. For instance, Boghossian once claimed that what follows

from the fact that you mean *green* by 'green' is a host of normative truths, such as that you ought to apply 'green' to something if and only if it is green (Boghossian, 1989). And Ginsborg (2012) claims not only that someone who means addition by 'plus' makes a certain pattern of normative judgments, but also that these judgments are those someone who means addition by 'plus' ought to make (see also Whiting, 2009). These philosophers seem to be saying that rules (requirements or principles) that determine meaning determine some normative facts.[11] It is *because* some course of action is required by meaning or content constitutive rules that one ought to, or has a normative reason to, carry out that course of action.

Unlike the view that meaning involves normative judgment, the view that meaning is a source of normativity does seem to have a direct bearing on the hard problem of intentionality. Suppose that semantic truths entail normative truths. If a satisfactory reductive explanation of meaning must explain everything that is entailed by the semantic truths, then any satisfactory reductive explanation of meaning will have to provide a reductive explanation of normativity as well. If normativity cannot be reductively explained, then meaning cannot be reductively explained either. Similarly, suppose that semantic facts metaphysically necessitate some normative facts. Then these normative facts supervene on the semantic facts, and any non-semantic supervenience base must metaphysically necessitate both the semantic and the normative facts. The claim that meaning is a source of normativity would have a bearing on whether meaning can be naturalized, if it were true. But as we shall see, a strong case can be made that it is not true.

Following Broome (2013), we can distinguish between a strong and a weak version of the claim that some norm *R* is a source of normativity:

> *Strong Normativity*: Necessarily, if *R* requires that you φ, then you ought to φ because *R* requires that you φ.
>
> *Weak Normativity*: Necessarily, if *R* requires that you φ, then the fact that *R* requires you to φ gives you a normative reason to φ.

In *Weak Normativity*, the reason *R* engenders could be merely *pro tanto*, that is, a reason that can be outweighed by other considerations. We have assumed, for the sake of argument, that meaning 'plus' is constituted by *Plus-Rule*. Is *Plus-Rule* either strongly or weakly normative?

Recall that *Plus-Rule* requires that you respond with the sum of any two numbers when asked. Of course, in many cases, responding with the sum is what you ought to do, but this is rarely simply because responding with the sum is required by the rule that constitutes what you mean. Sometimes, you ought to respond with the sum because you will benefit – if you are sitting an examination in mathematics, for instance. At other times, you ought to respond with the sum because this is what you morally ought to do. And sometimes you ought to respond with the sum because of a desire to communicate what you believe (cf. Hattiangadi, 2007; Glüer, 1999; 2001; Wikforss, 2001). But it is not necessarily the case that you ought to respond with the sum whenever you are asked, as in the evil demon case we have discussed previously.

Indeed, it is implausible that meaning constitutive rules such as these are even weakly normative. Of course, if you do have a normative reason to answer with the sum, this may accompany other prudential or moral reasons you might have to do so in a given context.

However, it is not plausible that the mere fact that answering with the sum is required by *Plus-Rule* necessarily gives you a genuinely normative reason to respond with the sum. One way to test this is to consider the evil demon case again, using the following heuristic: ask what a responsible agent ought to consider when deciding what to do. A responsible agent, when deciding what to do in a situation, will consider all of the facts that are normatively relevant to what she ought to do in that situation. Any *pro tanto* reason to do A or not to do A that is genuinely normative will be normatively relevant to whether an agent ought to do A. So, we can get a grip on which facts constitute *pro tanto* reasons for or against giving the answer '5' in the evil demon case by considering which facts will be taken into consideration when a responsible agent decides what to do in that case. Clearly, the fact that giving an answer other than '5' will cause the loss of life is normatively relevant, and constitutes a genuinely normative reason to give the answer '5.' An agent who failed to take this into account in deciding what to do would be irresponsible. In contrast, the fact that the agent is required by meaning-constituting rules to answer '125' does not seem to be normatively relevant in this case at all. An agent who failed to take this fact into account in deciding what to do, all things considered, would not be irresponsible. Thus it seems that the fact that *Plus-Rule* requires that you answer '125' does not even give you a *pro tanto* reason to do so in the evil demon case, as if that reason hung in the balance against the loss of life (Thomson, 2008).

Perhaps it will be tempting to respond by distinguishing semantic oughts from moral, prudential, or epistemic ones. In the case where you are faced with a choice between answering with the sum of two numbers and allowing thousands to die, or not answering with the sum and letting them live, there are two spheres of normativity that clash: what you *semantically* ought to do and what you *morally* ought to do. Even if what you ought to do, all things considered, is in this case identical with what you morally ought to do, it does not follow that it is not the case that you semantically ought to respond with the sum.

However, facts concerning what one *semantically* ought to do are merely norm-relative, not genuinely normative. We have assumed, for the sake of argument, that meaning is constituted by rules or requirements of some kind, that if you mean addition by 'plus' then you are semantically required to respond with the sum of any two numbers when asked. If what you semantically ought to do is just what you ought to do in view of the relevant semantic requirements, it is trivial that you semantically ought to do what is semantically required of you. Similarly, it is trivially true that you legally ought to do what the law requires of you, that you rationally ought to do what rationality requires of you, and that you morally ought to do what morality requires of you. Yet it remains an open question whether you ought to do what you morally ought to do, legally ought to do, or rationally ought to do. These are *qualified* oughts, which are norm-relative, not the *unqualified* ought, which is genuinely normative. Indeed, any system of normative requirements can be seen to give rise to qualified oughts. For instance, what you *satanically* ought to do is what you ought to do in view of the requirements of a satanic cult. In each of these cases, there is a *further question* whether one ought to do in the unqualified sense what one ought to do in some unqualified sense. In asking whether one ought to do what the law requires or whether one has reason to do what rationality requires, we use 'ought' in an *unqualified* sense (Broome, 2013).[12] It is this unqualified concept of 'ought' that figures in genuinely normative judgments. Correspondingly, there are unqualified concepts of permission, reason, and goodness.

Thus, even if we grant that you semantically ought to do what is semantically required of you, there is a further question whether these systems of norms are sources of normativity – whether you ought to or have reason to do what these systems of norms require of you in the unqualified sense of 'ought' and 'reason'. To be reminded that what you semantically ought to do is what is semantically required of you does not address this further question.

To sum up: when we ask whether morality, rationality, or semantics is a source of normativity, we are using the unqualified, genuinely normative concepts of 'ought' and 'reason.' As we have seen, it is implausible that meaning-determining rules are even weak sources of normativity. That is, it is implausible that if you mean addition by 'plus,' then the fact that you are semantically required to answer '125' gives you even a *pro tanto* reason to do so in the evil demon case. Though the view that meaning is a source of normativity would have a bearing on the hard problem of intentionality if it were true, it does not seem to be true.

4 The Normative Determination of Meaning

The view that meaning is a source of normativity is the view that semantic truths entail or metaphysically necessitate normative truths. An alternate version of the normativity thesis holds that the determination relation goes in the opposite direction. According to this view, the semantic or intentional facts are determined by normative facts (perhaps together with the natural facts). This view is motivated by the failure of attempts to give a naturalistic explanation of meaning. The normativist's proposal is to add normative truths to the reductive explanation or normative facts to the supervenience base. Even if meaning cannot be explained in *naturalistic* terms, perhaps it can be explained in naturalistic and normative terms.

This proposal has an obvious bearing on the hard problem of intentionality. For instance, one might hold that the semantic truths can be *reductively explained* in terms of the normative and natural truths. Second, one might hold that the semantic facts *supervene on* the natural and normative facts. Of course, one might hold both views.[13] It is important to note that none of these views holds much interest in the present context if meaning can be reductively explained in naturalistic terms alone. For then, any normative statements added to a reductive explanation would be nugatory. So, for the purposes of the present discussion, we can assume that meaning cannot be reductively explained in purely naturalistic terms. Our question is whether anything can be gained by adding normative statements to the reductive explanation or by adding normative facts to the supervenience base.

There are broadly speaking two approaches to the reductive explanation of meaning: holistic and atomistic. The holistic approach, associated most famously with Lewis (1974) and Davidson (1973), starts with the totality of relevant non-semantic information and then assigns beliefs, desires, and meanings in accordance with certain constraints on interpretation, such as most notably, some kind of principle of charity. The atomistic approach, associated with Fodor (1990), Dretske (1997), Millikan (1984), and others, starts by giving a reductive explanation of mental representations and then attempts to build a reductive explanation of beliefs, desires, and the meanings of sentences on this foundation. The contents of mental representations are reductively explained in terms of causal covariation: if a mental representation *X* causally covaries with the presence of *Fs*, then *X* picks out, or represents the *Fs*.

The atomistic approach is broadly dispositional. In its crudest version, it says that you mean addition by 'plus' if and only if you are disposed to answer with the sum of any two numbers when asked. Kripke argues that such a view faces the problem of error. Sometimes, you are disposed to make mistakes, and thereby fail to answer with the sum. What the dispositionalist needs is some way of distinguishing the meaning-constituting dispositions from those that are error-producing (Boghossian, 1989). This suggests a role for normativity in the reductive explanation of meaning. Perhaps it is a normative property of dispositions in virtue of which they are meaning-constituting rather than error-producing.

Which normative property? Perhaps the disposition to answer with the sum is meaning-constituting in virtue of being the disposition to give the answer that you ought to give. That is, you mean addition by 'plus' if and only if you are (a) typically disposed to answer with the sum, and (b) the sum is the answer that you ought to give. The trouble is that in the evil demon case it is false that you ought to answer with the sum. But it does not follow that you do not mean addition by 'plus.'

Alternatively, one might suggest that the disposition to answer with the sum is meaning-constituting iff it is a disposition to give the answer that is semantically correct. That is, you mean addition by 'plus' if and only if you are (a) typically disposed to answer with the sum, and (b) the sum is the answer that is semantically correct.[14] This might avoid the foregoing problem, but it violates a plausible circularity constraint on reductive explanation. The property of being semantically correct is itself a semantic property – it is the property of being correct *in virtue of meaning*. And an appeal to this property in a reductive explanation of meaning violates a circularity constraint on a satisfactory reductive explanation. If any semantic or properties figure in an explanation of intentionality, then that explanation is not fully reductive.

A third option might provide a middle way. Perhaps the meaning-constituting dispositions are the *rational* dispositions (cf. Wedgwood, 2009). A rational disposition is one that is triggered by the normative property of being rational. This might avoid the circularity worry, since what you rationally ought to do is what you ought to do *in virtue of the requirements of rationality*, not in virtue of meaning. However, it does not deal well with the evil demon case. Arguably, if you want most of all to avoid the loss of life, then giving the answer '5' is rationally required in that case. But if your disposition to answer '5' is rational, then according to the view currently under consideration, it follows that you mean schmaddition, rather than addition by 'plus.'

We can present the problem for the normative dispositionalist in the form of a dilemma. Either the dispositionalist makes use of the concept of *semantic* correctness in the reductive explanation of meaning, or she makes use of some other qualified or unqualified normative concepts. If she adopts the first strategy, then she violates a circularity constraint on reductive explanation. If she adopts the second strategy, then she gets the wrong result, such as that you mean schmaddition by 'plus.' Either way, reductive explanation fails.

Now let's return to the holistic approach. Consider Lewis's holistic defense of semantic naturalism. Lewis claims that a radical interpreter, who knows all of the physical truths *P* and grasps our concepts of belief, desire, and meaning, is in a position to deduce *a priori* what an arbitrary subject such as Karl believes, desires, and means. She achieves this feat by applying constraints on interpretation, such as a principle of charity: she represents Karl's beliefs and desires with probability and utility functions that assign the highest expected utility to the acts Karl performs, and assigns beliefs that are rational in light of Karl's evidence. Lewis's proposal is intended to be fully reductive, but he does not suggest that it is in

the slightest bit normative. Davidson, who also claims that radical interpretation is possible, is not after a reductive explanation of intentionality (Davidson, 1990), though he sometimes suggests that his account is normative (Davidson, 1990).[15] So, let us consider the hypothesis that radical interpretation is not possible if the radical interpreter's knowledge is restricted to the non-normative, naturalistic information, but that radical interpretation *is* possible if the radical interpreter is allowed to appeal to normative information of some kind. Can a judicious dose of normativity allow the interpretationist to avoid well-known difficulties that are otherwise unresolved?

One such unresolved difficulty is the 'permutation problem' faced by interpretationist accounts of the foundations of meaning (Davidson, 1973; Field, 1973; Putnam, 1980; Quine, 1964; Williams, 2007). Consider a radical interpreter who adopts a 'global descriptivist' strategy: her data consists of the set of sentences Karl holds true, or perhaps, would hold true under ideal conditions. This constitutes Karl's global theory, ***T***. She then assigns truth-conditions to sentences of Karl's language in such a way as to maximize truth in ***T***. The assignment of extensions to lexical items simply falls out of the assignments of meanings to sentences. The permutation problem for the global descriptivist goes as follows: from any interpretation, *x*, it is possible to construct a competing interpretation *y* that is equivalent to *x* in the assignment of truth-values to sentences of ***T***, but which assigns bizarre extensions to lexical items. To fix ideas, suppose that Karl holds true the sentence *s* 'all emeralds are green.' Now, let *x* be an interpretation that assigns the set of emeralds to 'emeralds' and the set of green things to 'green,' thereby rendering *s* true. We can easily construct an interpretation *y* that assigns the set of emerires to 'emeralds,' and the set of grue things to 'green,' where something is grue iff it is observed before *t* and found to be green, or is blue otherwise, and where something is an emerire iff it is observed before *t* and found to be an emerald, or is a sapphire otherwise. Though *y* is a bizarre interpretation, it too renders *s* true.[16]

Can this problem be resolved if the radical interpreter is allowed recourse to some normative information? It is sometimes suggested that radical interpretation will succeed if the radical interpreter is allowed information about what is *rational*. That is, if the radical interpreter knows which beliefs are rational in light of Karl's evidence, or which belief/desire pairs would rationalize Karl's behavior, then she will be able to apply the principle of charity and thereby rule out deviant interpretations. One might argue that Karl's evidence confirms the belief that all emeralds are green, while it does not confirm the hypothesis that all emerires are grue (Goodman, 1983; Lewis, 1984). As Goodman would put it, neither 'emerire' nor 'grue' are instance confirmable, projectible predicates. However, as Davidson argued, only sets of predicates can jointly be deemed projectible and instance confirmable. Just as an observation of a green emerald confirms the hypothesis that all emeralds are green, an observation of a grue emerire confirms the hypothesis that all emerires are grue (Davidson, 1980, p. 226). Which beliefs are rationalized by Karl's evidence will depend in part on how Karl conceptualizes the evidence, which perceptual beliefs his perceptual experiences generate. If he acquires the perceptual belief that there is a grue emerire before him, then it is rational for him to increase his confidence in the hypothesis that all emerires are grue. So, even if the radical interpreter knows which beliefs would be rational in the face of the evidence, since she doesn't know how the evidence is conceived, this information will not help her to rule out the permuted interpretation.

Another thought might be to allow the radical interpreter information about which of Karl's utterances are semantically correct, or which ones Karl semantically ought to make. Once again, this suggestion violates the circularity requirement on reductive explanation. If

'semantically correct' means 'required by the semantic requirements' and the semantic requirements are constitutive rules like *Plus-Rule*, statements concerning semantic correctness are equivalent to semantic statements. Thus, the holist faces a dilemma very similar to that of the atomist. If the holist appeals to facts about semantic correctness, she violates a circularity requirement, whereas if she appeals to facts of some other kind, such as facts about what is rational, she fails to provide a reductive explanation.

Thus it is not obvious that there is anything to be gained by adding normative truths to the reductive explanation of meaning. At any rate, doing so does not seem to resolve well-known difficulties that beset such reductive explanation. As a consequence, we have no positive reason to think that the semantic facts are determined by normative facts.

5 The Normativity of Semantic Concepts

In his recent book, Allan Gibbard (2012) argues not that meaning is normative, but that the concept MEANING is normative, much like paradigmatic normative concepts such as GOOD and OUGHT. He claims further that meta-representational ascriptions of meaning, such as "Shah means addition by 'plus'" are normative, not descriptive, as are self-ascriptions of meaning, and ascriptions of propositional attitudes. Gibbard combines this view with his own brand of expressivism about normativity according to which assertions of normative statements do not purport to describe how things are, but express something akin to plans. Similarly, he maintains that assertions of metalinguistic ascriptions of meaning, such as 'Shah means addition by "plus"' do not describe, but express plans. Gibbard's claims here appear to be consonant with Kripke's 'skeptical solution' to the skeptical problem.[17] Yet, his approach is further developed. He aims to shed light on Kripke's problem by exploiting an analogy with metaethics. In metaethics, the expressivist says that because moral sentences are non-descriptive, there is no need to postulate any moral facts out there in the world to correspond to them. By the same token, Gibbard suggests that if semantic ascriptions are non-descriptive, there is no need to postulate any semantic facts out there in the world to correspond to them. He hopes thereby to dissolve Kripke's problem.

Gibbard introduces his claim that the concept MEANING is normative with the slogan "'means' entails 'ought.'" "Ascriptions of meaning," he says, "imply straight ought ascriptions" (Gibbard, 2012, p. 11). He adds that these entailments "could only be analytic or conceptual" (Gibbard, 2012, p. 23). That is, ascriptions of meaning, such as "Pierre means DOG by 'chien,'" entail ascriptions of oughts, such as "Pierre ought to…" and these entailments hold in virtue of the meaning of 'meaning' or equivalently, in virtue of the concept MEANING. Anyone who grasps the concept MEANING must accept the normative entailments of meaning ascriptions.

What are the normative entailments of meaning ascriptions? Here are two examples drawn from Gibbard's discussion, where the first statement of each pair is said to conceptually or analytically entail the second:[18]

(1a) The sentence 'Schnee ist weiss' in Ursula's language means SNOW IS WHITE.
(1b) Ursula ought to accept 'Schnee ist weiss' iff snow is white.

(2a) Pierre's sentence 'les chiens aboient' means DOGS BARK.
(2b) If Pierre has sufficient undefeated evidence that dogs bark in an epistemic circumstance *E*, then Pierre ought to accept the sentence 'les chiens aboient' in *E*.

If these entailments are analytic, as Gibbard suggests, then anyone who grasps the concept MEANING must accept that (1a) entails (1b), and that (2a) entails (2b), on pain of irrationality or conceptual confusion.[19] However, it is possible for someone to sensibly deny these entailments without thereby displaying either irrationality or confusion about the concept MEANING.

First, consider an evidentialist, such as William Clifford, who holds that one ought never to believe anything on insufficient evidence (Clifford, 1877). Suppose that we present Clifford with a hypothetical scenario in which Ursula is a speaker of German; the sentence 'Schnee ist weiss' means that snow is white in Ursula's language; snow is in fact white; but Ursula has never seen snow, and lacks any testimonial evidence as to its color. Since Ursula lacks sufficient evidence that snow is white, Clifford would judge that Ursula ought not to believe SNOW IS WHITE. Moreover, note that according to Gibbard, one way to believe SNOW IS WHITE is to accept a sentence in a language one understands that means that snow is white (cf. Gibbard, 2012, pp. 27, 44). It follows from this that Clifford's judgment that Ursula ought not to believe SNOW IS WHITE is tantamount to the judgment that Ursula ought not to accept the sentence 'Schnee ist weiss.' Thus, Clifford would accept (1a) but not (1b). Yet, in so doing, he would display neither irrationality nor confusion about the concept MEANING. Even if we think Clifford is mistaken, his mistake is not conceptual, because his reasons for rejecting the entailment from (1a) to (1b) have to do with his acceptance of evidentialism, and nothing to do with his grasp of the concept MEANING. Thus, (1a) does not analytically entail (1b) (cf. Whiting, 2015).

It can be shown in a similar fashion that the entailment from (2a) to (2b) is not analytic. Consider Blaise Pascal (1670), who holds that there can be pragmatic reasons for belief. Suppose that we present Pascal with a hypothetical situation in which Pierre is a monolingual speaker of French who has sufficient undefeated evidence that dogs bark, but is promised eternal bliss if he does not accept the sentence 'les chiens aboient' and eternal torture if he does. With regard to this situation, Pascal would accept (2a), but deny that Pierre ought to believe DOGS BARK. Since Gibbard assumes that believing DOGS BARK is tantamount to accepting a sentence in one's language that means DOGS BARK, it follows that Pascal would reject (2b). Yet in so doing, he need display neither irrationality nor confusion about the concept MEANING. Even if Pascal's view that there can be pragmatic reasons for belief is mistaken, his mistake is not conceptual. His grounds for rejecting the inference from (2a) to (2b) have to do with his pragmatism, not with his grasp of the concept MEANING.[20] Thus, (2a) does not analytically entail (2b).

One way to respond to these objections might be to appeal to the notion that there are distinct spheres of normativity: such as *inter alia*, epistemic, prudential, and semantic normativity. One might then say that what (1a) and (2a) entail are (1c) and (2c), respectively:

(1c) Ursula *semantically* ought to accept 'schnee ist weiss' iff snow is white.
(2c) If Pierre has sufficient undefeated evidence that dogs bark in an epistemic circumstance *E*, then Pierre *semantically* ought to accept the sentence 'les chiens aboient' in *E*.

One could argue that even if Clifford could sensibly reject the inference from (1a) to (1b), he would have to be conceptually confused to reject the inference from (1a) to (1c). What Clifford ought to say about Ursula's case is that she *epistemically* ought not to believe that snow is white, but that she *semantically* ought to accept the sentence 'schnee ist weiss' nonetheless. Similarly, one could argue that even if Pascal could sensibly reject the inference from (2a) to (2b), he would have to be conceptually confused to reject the inference from

(2a) to (2c). What Pascal should say is that while Pierre *prudentially* ought not to believe DOGS BARK, he *semantically* ought to accept the sentence 'les chiens aboient' nonetheless.

However, this response is not available to Gibbard, who claims that semantic ascriptions entail 'Ewing's primitive oughts.' In Ewing's sense of 'ought' to say 'you ought to do *X*' is to say that you ought to do *X*, all things considered, or that you have a conclusive reason to do *X* (Gibbard, 2012, p. 14). Ewing's sense of 'ought' is just the unqualified sense of 'ought' that we have already come across. Thus, Gibbard's official view is that (1a) entails (1b), and (2a) entails (2b), where both (1b) and (2b) are understood as all things considered oughts, not that (1a) entails (1c) nor that (2a) entails (2c).

Moreover, it *is* possible for someone to sensibly reject the inference from (1a) to (1c), or from (2a) to (2c), without displaying any conceptual confusion, so even these entailments are not plausibly analytic. To see why, it will help to generalize (1c) and (2c). Let *S* be a subject; assume that sentence 's' means *P* in the language spoken by *S*, and that 's' is true iff *p*. We can generalize (1c) and (2c) as follows:

(1d) *S semantically* ought to accept 's' iff *p*.
(2d) If *S* has sufficient undefeated evidence that *p* is true in an epistemic circumstance *E*, then *S semantically* ought to accept 's' in *E*.

First, note that (1d) and (2d) can come into conflict.[21] Assume that Anna speaks Swedish, in which 'häxor existerar' means WITCHES EXIST; it is not the case that witches exist, yet Anna has sufficient undefeated evidence that they do. In this case, (1d) will entail that it is not the case that Anna ought to accept 'häxor existerar' whereas (2d) will entail that *S* ought to accept it. A proponent of a truth-conditional meaning theory, who holds that truth is all that semantically matters for sentence acceptance (belief), would accept (1d) but not (2d), whereas a verificationist, who holds that evidence is all that matters semantically for sentence acceptance (belief), would accept (2d) but not (1d). Yet neither of them need be confused about the concept MEANING. Their dispute concerns the nature of the semantic norms. It is possible for someone to endorse either position without thereby displaying confusion about the concept MEANING.

Perhaps Gibbard would argue that the dispute between the advocate of (1d) and the advocate of (2d) is a substantive normative dispute. He might suggest that the concept MEANING is normative in the sense that to grasp the concept MEANING one must accept that meaning ascriptions have *some* normative implications – one ought to accept *either* (1d), (2d), or similar – though it does not matter *which* normative implications one takes semantic ascriptions to have. This would suggest that Gibbard's view is that the concept MEANING is a 'thin' normative concept like GOOD or OUGHT which lacks any significant descriptive content, rather than a 'thick' normative concept like JUSTICE or COURAGE, which has a significant descriptive content (cf. Gibbard, 2012, p. 39).

Why should we think that MEANING is a thin normative concept? Many semantic anti-normativists will be inclined to deny that meaning ascriptions have any normative entailments. If this denial is sensible, then MEANING is not a thin normative concept. However, since I count myself among the semantic anti-normativists, I am inclined to think that our reasons for denying meaning ascriptions have any normative entailments are sensible.[22] But, what positive reasons does Gibbard give for accepting the normativity thesis?[23] The chief reason is that the normativity of the concept MEANING explains why certain basic oughts follow from semantic ascriptions 'invariably' (Gibbard, 2012, p. 16). However, this

rings hollow in light of the foregoing discussion. As we have seen in the cases above, there are sensible grounds to question whether semantic ascriptions really do entail basic oughts invariably.

Another reason Gibbard provides is that viewing the concept MEANING as normative promises a satisfactory expressivist resolution to Kripke's problem. Whether this is a good reason to accept that MEANING is a thin normative concept depends on whether expressivism does offer a satisfactory resolution to Kripke's problem.

So, does it? To assess this, it will help to have a picture of how the expressivist solution goes in the moral domain. In this domain, the expressivist says that moral statements do not describe but prescribe, and that moral judgments are not straightforwardly factual beliefs. Rather, moral statements express non-cognitive attitudes of some kind – in Gibbard's view, they express a special kind of contingency plan. If moral statements express plans, then there is no need to postulate any moral properties or moral facts out there in the world to correspond to them. In this way, the expressivist proposes to dissolve the hard problem of morality. If there is no need to postulate a property of goodness, there is no need to explain, in non-moral terms, what goodness consists in. As Gibbard puts it, the expressivist gives an *oblique* naturalistic explanation of morality. He rejects *straight* reductive naturalist attempts to reductively explain what makes it the case that something has the moral properties that it does, but explains why no such straight reductive naturalistic explanation of morality is needed. The expressivist can show that morality is in some sense continuous with natural science while rejecting any reductive explanation of morality. Gibbard argues that all that is needed for an oblique naturalistic explanation of morality is an expressivist treatment of language and thought about morality.

However, this style of explanation breaks down when applied to meaning. First of all, note that its starting point is the *semantic* thesis that meaning ascriptions do not describe, but express plans. This thesis concerns the *meaning* of a certain class of sentences – the semantic ascriptions. This thesis appears to commit the expressivist at least to the existence of one semantic property: the property of expressing a plan. A straight naturalist would attempt to give a reductive explanation of this semantic property, to explain what this semantic property consists in. But as an expressivist, Gibbard is committed to giving an oblique explanation of this semantic property which involves explaining why there is no need to postulate such a property in the first place. Gibbard might try saying that the expressivist's own claim that semantic statements express plans is itself an expression of a plan. But there are two difficulties with this move. First, it seems to give rise to a vicious regress of oblique explanation. Since the first-order claim that semantic ascriptions express plans committed the expressivist to the existence of at least one semantic property, the metalinguistic claim that the expressivist's first-order claim expresses a plan also commits him to the existence of a semantic property, and there is no explanation for why a straight reductive explanation of what this property consists in is put off indefinitely (see Hattiangadi, 2015).

Second, if the expressivist's central claim is indeed true, and all semantic statements express plans, then the expressivist's own claim that semantic statements express plans itself expresses a plan. But it is hard to see what bearing this has on the hard problem of intentionality. For, even if the expressivist expresses the plan to treat semantic statements as expressing plans, this does not entail that semantic statements do express plans. Consider the analogy again with morality. In the moral case, the expressivist's claim that moral statements express plans is naturally understood to describe the semantics of moral statements.

Given that this semantics is descriptively accurate, there is no need to postulate any moral properties to correspond to moral statements. In the semantic case, the expressivist must be understood to express a plan in claiming that semantic ascriptions express plans. But it does not follow from the expression of this plan that there is no need to postulate any moral properties (see Hattiangadi, 2015).[24]

So, expressivism does not offer a route to a satisfactory solution to Kripke's problem. This undermines Gibbard's central argument for the view that the concept MEANING is normative, which we already had independent reasons to reject.

6 Conclusion

In sum, we have seen that the normativity of meaning really has very little bearing on the hard problem of intentionality in any of the four senses given above. The view that meaning involves rule-following or a normative judgment of some kind is untenable, and in any case, has no bearing on the hard problem of intentionality. Both the hypothesis that meaning is a source of normativity and the hypothesis that meaning is determined by normativity might in principle have a bearing on the hard problem of intentionality. However, the view that meaning is a source of normativity is implausible, and there is little evidence that known difficulties with the reductive analysis of meaning can be resolved by adding normativity to the explanation. Finally, Gibbard's recent suggestion that the concept MEANING is normative is implausible and his proposed expressivist resolution of the hard problem of intentionality appears to be untenable.

Notes

1 I will adopt the convention of using small capital letters to denote concepts and thoughts.

2 This is what Kathrin Glüer and Åsa Wikforss call 'ME normativism,' or 'meaning-engendered' normativism.

3 I assume that to accept a sentence is either to assert it, assent to it, or to be disposed to do so.

4 This is much like the 'hard problem of consciousness' discussed by Chalmers (1996).

5 Though there are many ways to make supervenience more precise, I will assume that to say that the semantic facts supervene on the non-semantic facts is to say that any metaphysically possible world that is a minimal non-semantic duplicate of our world, is a semantic duplicate of our world, where a minimal non-semantic duplicate of our world is a world that satisfies all of the non-semantic truths and contains no 'extras' (cf. Jackson, 1998).

6 See Hattiangadi (2007), Finlay (2010). For a formal semantic account of norm-relative statements, specifically containing 'must' or 'can,' see Kratzer (1977; 1981). In Kratzer's semantics, norm-relative statements are treated as relational modal claims. They are statements of metaphysical necessity restricted in relation to two parameters: a modal base, W, the set of worlds consistent with everything that is believed as common ground in the context, and an ordering $\geq_o$ on W, the source of which, O, is in this case your desires, also made salient by the context. Thus, she says, 'it must be that p' is true iff p is true at every highest ranked world $w \in W$ (as ranked by O), 'it may be that p' is true iff p is true at at least one of the highest ranked worlds.

7 I use 'fact' and 'truth' interchangeably.

8 Thanks to Marianna Bergamaschi Ganapini for this suggestion.

9 It has been common in the discussion of the normativity of meaning to distinguish hypothetical or instrumental oughts from categorical oughts, and identify genuine normativity with the latter

(cf. Hattiangadi, 2006; 2007; Glüer, 1999; Glüer and Wikforss, 2010; Wikforss, 2001). Though it is clear that instrumental oughts are a class of norm-relational oughts, it is not entirely clear where categorical oughts fit in. The concept of a categorical ought goes back to Kant's categorical imperative, which itself is an absolute, unconditional requirement. However, if what you categorically ought to do is what you ought to do in view of a Kantian categorical imperative, then categorical oughts are norm-relational in my sense. In the contemporary discussion of the normativity of meaning, categorical oughts are often characterized simply as non-instrumental. But the class of norm-relative oughts is larger than the class of instrumental oughts, since some non-instrumental oughts – such as role-oughts and functional oughts – are non-instrumental but nevertheless norm-relational.

10 Anti-normativists have argued that the fact that I mean addition by 'plus' may be relevant to what I ought to do, conditional on the desire to speak truthfully (and other relevant background conditions), but this does not make this fact normative. By the same token, the fact that it is raining is relevant to whether I ought to take an umbrella when I go out, conditional on the desire to remain dry (together with relevant background conditions). And this does not make facts about the weather normative (Glüer, 1999; Hattiangadi, 2006; 2007; Wikforss, 2001). Verheggen's response to this objection is to argue that there is a disanalogy: instrumental semantic oughts are essential to meaning, whereas instrumental oughts about the weather are not. If it is false that if I want to tell the truth, then I ought to say '58 + 67 = 125,' then I do not mean addition by 'plus,' whereas if it is false that if I want to avoid getting wet, then I ought to take an umbrella, nothing follows about the weather. Yet, these cases seem to be parallel: if it is not the case that it is necessary to carry an umbrella in order to avoid getting wet, then holding other relevant background conditions fixed (I plan to go out of doors, I don't have a raincoat, rain makes you wet, umbrellas protect from the rain, etc.), it does follow that it is not raining. Furthermore, Verheggen's point does not speak to the objection: even if instrumental oughts are essential to meaning, they are not genuinely normative.

11 This is what Glüer and Wikforss (2010) call ME normativism.

12 Sometimes, norm-relational oughts are contrasted with the *all things considered* ought. But the concept of the all things considered ought seems to be norm-relational, since what you ought to do all things considered is what you ought to do, *in view of everything that is normatively relevant*. So, the concept of the all things considered ought is distinct from the concept of the unqualified ought. Nevertheless, it is plausible that the two concepts are co-extensive: necessarily, you ought to do A iff you ought to do A, all things considered.

13 A twist on this is the view that semantic truths can be reductively explained in terms of the normative and natural truths, but that the semantic facts supervene on the natural facts alone. We will consider one version of this view in the next section: Gibbard holds that semantic concepts are normative, but that they pick out natural properties.

14 Of course, it seems as though you are not disposed to respond with the sum in the evil demon case. On a very crude dispositional view, it would follow from this that you do not mean addition by 'plus.' However, an alternative explanation of the evil demon case is available. A more sophisticated dispositionalist might say that though you are disposed to respond with the sum whenever asked, this disposition is 'blocked' or 'masked' by the disposition to do what you think is morally required in the evil demon case.

15 It is controversial whether Davidson truly is a normativist (cf. Engel, 2001, and Glüer, 2001).

16 It might be tempting to defend Lewis's appeal to 'naturalness' as a solution to the permutation problem (cf. Lewis, 1983; 1984). However, we have assumed, for the sake of argument, that the permutation problem cannot be solved by recourse to the non-normative information alone. So we can set this kind of approach aside. Our question is whether the permutation problem can be solved if the radical interpreter is allowed recourse to some normative information.

17 Thanks to Alex Miller for pointing this out.

18 The first example occurs on p. 40 of Gibbard (2012), with minor stylistic variations, while the second is reconstructed from Gibbard's discussion of analyticity and synonomy in chapter 6. Some further examples occurring in the text are somewhat puzzling. For instance, Gibbard says: "I ought not to believe, all at once, that snow is white and that nothing is white – and that ties in with the meaning of our term 'nothing'" (Gibbard, 2012, p. 13). Even if it is true that I ought not to believe both that snow is white and that nothing is white, this seems not to have anything to do with the meaning of the term 'nothing.' For it is arguably true that Kaveri, a monolingual speaker of Konkani, ought not to believe both that snow is white and that nothing is white, yet it is difficult to see what this might have to do with the meaning of the English term 'nothing.'

19 Gibbard's claim is that the concept MEANING is normative, and this is how he suggests we might capture the normativity of this concept. Note that his view is not that all concepts, such as SNOW, WHITE, and so forth are normative. That is, it is not, according to him, grasp of these concepts that entails acceptance of the normative entailments involving them.

20 One might try to argue that he fails to grasp the concept of belief, because beliefs cannot be formed at will, but that would not help Gibbard in his efforts to argue that the concept of meaning is normative. In any case, Gibbard accepts that beliefs can be formed at will (Gibbard, 2012).

21 It is worth noting that Gibbard distinguishes between subjective and objective oughts: what you objectively ought to do is what you ought to do in light of everything that is the case, what you subjectively ought to do is constrained in some way by your state of information. He claims that semantic oughts are subjective oughts, which suggests that he would ultimately reject the implication from (1a) to (1b). He can thereby avoid the inconsistency between the two sets of analytic entailments. But he cannot thereby avoid the objections presented above, since those turn on the thought that one might sensibly reject either entailment without displaying conceptual confusion.

22 See, e.g., Glüer (1999), Glüer and Wikforss (2009), Hattiangadi (2006; 2009; 2007), Wikforss (2001). For a comprehensive review of the issue, see Glüer and Wikforss (2010).

23 Gibbard (2012, p. 16) distinguishes between a weak normativity thesis, according to which meaning ascriptions have normative entailments, and a strong normativity thesis, according to which the meaning of an expression is defined by the pattern of oughts it entails. Though Gibbard purports to defend both theses, it is only the weak normativity thesis that is relevant to the present discussion.

24 Of course, a cognitivist realist about normativity might claim that the concept of meaning is normative while rejecting an expressivist treatment of semantic normativity. So, the view that the concept of meaning is normative does not stand or fall with expressivism. However, a cognitivist realist about normativity who claims that the concept of meaning is normative would be committed to the view either that meaning is a source of normativity or that semantic facts are determined by normative (and natural) facts. Since I consider and reject those views separately, I set them aside in this section.

References

Bilgrami, A. 1992. *Belief and Meaning*. Oxford: Blackwell.

Block, N. 1986. "Advertisement for a semantics for psychology." *Midwest Studies in Philosophy*, 10(1): 615–678.

Boghossian, P. 1989. "The rule-following considerations." *Mind*, 98(392): 507–549.

Boghossian, P. 2003. "The normativity of content." *Philosophical Issues*, 13(1): 31–45.

Boghossian, P. 2005. "Rules, meaning and intention: discussion." *Philosophical Studies*, 124(2): 185–197.

Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Broome, J. 2013. *Rationality Through Reasoning*. Oxford: Oxford University Press.

Bykvist, K., and A. Hattiangadi. 2013. "Belief, truth and blindspots." In *The Aim of Belief*, edited by T. Chan, pp. 100–122. Oxford: Oxford University Press.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Clifford, W. K. 1877. "The ethics of belief." In *Contemporary Review*. Reprinted in *The Ethics of Belief and Other Essays*, edited by T. Madigan, pp. 70–96. Amherst, MA: Prometheus, 1999.

Davidson, D. 1973. "Radical interpretation." *Dialectica*, 27(3–4): 313–328.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. 1990. "The structure and content of truth." *Journal of Philosophy*, 87(6): 279–328.

Dretske, F. 1997. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Engel, P. 2000. "Wherein lies the normative dimension in meaning and mental content?" *Philosophical Studies*, 100(3): 305–321.

Engel, P. 2001. "Is truth a norm?" In Kotatko and Pagin, 2001, pp. 37–51.

Field, H. 1973. "Theory change and the indeterminacy of reference." *Journal of Philosophy*, 70(14): 462–481.

Field, H. 1977. "Logic, meaning and conceptual role." *Journal of Philosophy*, 74(7): 379–409.

Finlay, S. 2010. "Recent work on normativity." *Analysis*, 70(2): 331–346.

Fodor, J. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Gampel, E. H. 1997. "The normativity of meaning." *Philosophical Studies*, 86(3): 221–242.

Gibbard, A. 1994. "Meaning and normativity." *Philosophical Issues*, 5: 95–115.

Gibbard, A. 2003. "Thoughts and norms." *Philosophical Issues*, 13(1): 83–98.

Gibbard, A. 2012. *Meaning and Normativity*. Oxford: Oxford University Press.

Ginsborg, H. 2011. "Primitive normativity and skepticism about rules." *The Journal of Philosophy*, 108(5): 227–254.

Ginsborg, H. 2012. "Meaning, understanding and normativity." *Aristotelian Society*, suppl. vol. 86(1): 127–146.

Glock, H.-J. 1996. "Necessity and normativity." In *The Cambridge Companion to Wittgenstein*, edited by H. Sluga and D. G. Stern, pp. 198–225. Cambridge: Cambridge University Press.

Glock, H.-J. 2005. "The normativity of meaning made simple." In *Philosophy–Science–Scientific Philosophy*, edited by A. Beckermann and C. Nimtz, pp. 219–241. Paderborn, Germany: Mentis.

Glüer, K. 1999. "Sense and prescriptivity." *Acta Analytica*, 14: 111–128.

Glüer, K. 2001. "Dreams and nightmares: conventions, norms and meaning in Davidson's philosophy of language." In Kotatko and Pagin, 2001, pp. 53–74.

Glüer, K., and P. Pagin. 1999. "Rules of meaning and practical reasoning." *Synthese*, 117(2): 207–227.

Glüer, K., and Å. Wikforss. 2009. "Against content normativity." *Mind*, 118(469): 31–70.

Glüer, K., and Å. Wikforss. 2010. "The normativity of meaning and content." *Stanford Encyclopedia of Philosophy*, winter edn, edited by E. N. Zalta. http://plato.stanford.edu/archives/win2010/entries/meaning-normativity/ (accessed August 24, 2016).

Goodman, N. 1983. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Harman, G. 1999. *Reasoning, Meaning and Mind*. Oxford: Oxford University Press.

Hattiangadi, A. 2003. "Making it implicit: Brandom on rule following." *Philosophy and Phenomenological Research*, 66(2): 419–431.

Hattiangadi, A. 2006. "Is meaning normative?" *Mind & Language*, 21(2): 220–240.

Hattiangadi, A. 2007. *Oughts and Thoughts: Rule-Following and the Normativity of Content*. Oxford: Oxford University Press.

Hattiangadi, A. 2009. "Semantic normativity in context." In *New Waves in the Philosophy of Language*, edited by S. Sawyer, pp. 87–107. London: Palgrave MacMillan.

Hattiangadi, A. 2015. "The limits of expressivism." In *Meaning Without Representation: Essays on Truth, Expression, Normativity, and Naturalism*, edited by S. Gross, N. Tebben, and M. Williams, pp. 224–242. Oxford: Oxford University Press.

Ichikawa, J., and B. Jarvis. 2013. *The Rules of Thought*. Oxford: Oxford University Press.

Jackson, F. 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Kotatko, P., and P. Pagin, eds. 2001. *Interpreting Davidson*. Stanford, CA: Center for the Study of Language and Information.

Kratzer, A. 1977. "What 'must' and 'can' must and can mean." *Linguistics and Philosophy*, 1(3): 337–355.

Kratzer, A. 1981. "The notional category of modality." In *Words, Worlds, and Contexts*, edited by H.-J. Eikmeyer and H. Rieser, pp. 38–74. Berlin: De Gruyter.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: MIT Press.

Kusch, M. 2006. *A Sceptical Guide to Meaning and Rules. Defending Kripke's Wittgenstein*. Chesham, UK: Acumen.

Lance, M., and J. O'Leary-Hawthorne. 1998. *The Grammar of Meaning: Normativity and Semantic Content*. New York: Cambridge University Press.

Lewis, D. 1974. "Radical interpretation." *Synthese*, 23: 331–344.

Lewis, D. 1983. "New work for a theory of universals." *Australasian Journal of Philosophy*, 61(4): 343–377.

Lewis, D. 1984. "Putnam's paradox." *Australasian Journal of Philosophy*, 62(3): 221–236.

McDowell, J. 1984. "Wittgenstein on following a rule." *Synthese*, 58(3): 325–363.

McHugh, C. 2014. "Fitting belief." *Proceedings of the Aristotelian Society*, 114(2): 167–187.

Millar, A. 2002. "The normativity of meaning." *Royal Institute of Philosophy Supplement*, 51: 57–73.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

Pascal, B. 1670. *Pensées*. Translated by W. F. Trotter. London: Dent, 1910.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: Harvard University Press.

Putnam, H. 1980. "Models and reality." *Journal of Symbolic Logic*, 45(3): 464–482.

Quine, W. V. O. 1964. *Word and Object*. Cambridge, MA: MIT Press.

Soames, S. 1997. "Skepticism about meaning: indeterminacy, normativity and the rule-following paradox." *Canadian Journal of Philosophy*, suppl. vol. 23: 211–249.

Thomson, J. J. 2008. *Normativity*. Peru, IL: Open Court.

Verheggen, C. 2011. "Semantic normativity and naturalism." *Logique et Analyse*, 216: 553–567.

Wedgwood, R. 2009. "The normativity of the intentional." *The Oxford Handbook of the Philosophy of Mind*, edited by B. P. McLaughlin and A. Bekermann, pp. 421–436. Oxford: Clarendon Press.

Whiting, D. 2007. "The normativity of meaning defended." *Analysis*, 67(2): 133–140.

Whiting, D. 2009. "Is meaning fraught with ought?" *Pacific Philosophical Quarterly*, 90(4): 575–599.

Whiting, D. 2015. "Review of *Meaning and Normativity* by Allan Gibbard." *European Journal of Philosophy*, 23(S1): E14–E18.

Wikforss, Å. 2001. "Semantic normativity." *Philosophical Studies*, 102(2): 203–226.

Williams, J. R. G. 2007. "Eligibility and inscrutability." *Philosophical Review*: 116(3): 361–399.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Zalabardo, J. L. 2012. "Semantic normativity and naturalism." In *Continuum Companion to Philosophy of Language*, edited by M. G. Carpintero and M. Kölbel, pp. 203–227. London: Continuum.

26

Indeterminacy of Translation

CRISPIN WRIGHT

W. V. O. Quine's contention that translation is indeterminate has been among the most widely discussed and controversial theses in modern analytical philosophy. It is a standard-bearer for one of the late twentieth century's most characteristic philosophical preoccupations: the skepticism about semantic notions which is also developed in Kripke's interpretation of Wittgenstein on rules (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING) and which many have read into Putnam's 'model-theoretic' assault on realism (see Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM). The more general concern reflected by these arguments is how space can be found for the reality of meanings – and indeed for other norms, like the ethical – in a world whose fundamentals, as the orthodox wisdom has it, are apt for complete characterization by the methods and vocabulary of physical science.[1]

If Quine's arguments succeed, this is not a concern which we should seek to allay, since there can be no satisfactory answer to it. At least, that will be the position if, with Quine, we take it that if translation is indeterminate, so is meaning itself, so that there are accordingly no genuine facts about meanings for a satisfactory world-view to accommodate. Quine envisages, moreover, that a consequential indeterminacy must spread outwards to infect ordinary "folk" or intentional psychology which comprises states, like beliefs and desires, which are in part identified by their content, as well as modal properties of statements, such as necessity and possibility, which functionally depend upon those statements' meanings (see Chapter 31, MODALITY). Much is at stake, then, in Quine's thesis.

The discussion to follow is organized as follows. §§1 and 2 will offer some initial reflections on the content and implications of the indeterminacy thesis, and of the presuppositions that Quine makes in treating it as a stepping-stone to semantic irrealism. §3 then distinguishes Quine's two principal arguments[2] for the thesis: the famous 'gavagai' argument ('from Below') of *Word and Object* (1960), and the argument ('from Above') from the underdetermination of empirical theory by data emphasized in "On the reasons for the indeterminacy of translation" (1970), and lays out the essentials of the former.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

§4 appraises this argument in the light of Gareth Evans's discussion in his "Identity and predication" (1975). §5 assesses the cogency of Evans's objections. §6 turns to the second and more radical argument, laying out certain basic distinctions and implications; and §7 is concerned with its appraisal.

Quine's contribution to these issues is, in effect, no less than to have invented them and to have set the agenda for all subsequent discussion. He has continued a vigorous engagement in that discussion, and students of some of his more recent contributions, for instance in *Pursuit of Truth* (1990) or the "Three indeterminacies" paper (1989), will have noted significant developments and changes of emphasis. But these are beyond the purview of this chapter, whose concern is not with scholarship of the perhaps revisionist tendencies of Quine's more recent thought, but simply with the structure, power, and philosophical background of the classic Quinean arguments.

1 What Does the Indeterminacy of Translation Involve?

It's commonplace that expressions in one language may resist a fully satisfactory translation into another: that there is, for instance, no exact English equivalent, capturing all its existentialist overtones, of the French *ennui* (a kind of jaded detachment), or the special piquancy of the German *Schadenfreude* (a form of pleasurable excitement at another's misfortune). This commonplace has little to do with Quine's thesis. Quine's claim is not that exact translation is sometimes impossible, but that *there is no such thing as* exact translation: that for any expression, in any language, there will inevitably be a range of alternative translations of it into any particular language each of which, in conjunction with coordinating adjustments in the translation of other expressions, will equally well – and *unimprovably* – accommodate all the behavioral data concerning speakers' use of the translated language.[3]

Quine argues his case for this thesis in the context of *radical* translation. Radical translation is the translation of a hitherto wholly untranslated language, about whose syntax, semantics, and etymology we are in a position to make no prior assumptions whatever. All we are assumed to have to go on is our observation of the use of the language. More specifically: we are assumed to be able to identify behavior on the part of its native speakers which constitutes assent to and dissent from particular utterances in the language, and we are assumed to be able to observe the circumstances in which such assent or dissent takes place. We are also allowed to suppose that we are able to interact with native speakers in particular contexts, to put utterances to them in their own language for assent or dissent, and in general to encourage the production of evidential data for our translation, rather than merely passively observe. Quine's claim is that if a project of translation is undertaken under these circumstances, then there are bound to be intuitively incompatible claims about the meanings of (what we identify as) expressions in the natives' language such that, no matter how extensive the data which we proceed to gather, it will in principle never give us a reason to prefer one such claim to another.

Why, it may be wondered, the focus on radical translation? Why is the situation of someone engaged in so unusual a project, on so impoverished a basis of collateral information, of particular interest? Well, consider how it might be that radical translation was indeterminate, yet *non-radical* translation in certain cases – say the translation of some parts of French into English – was a fully determinate matter. There's no doubt that in

translating a French utterance into English, we will make all kinds of assumptions – about the accuracy of dictionaries, the context, the purposes of the speaker, and so on – which effectively may uniquely determine the translation of a particular word in it. But what *justifies* these assumptions? Quine's thought is that their justification would ultimately have to come back to what could be appreciated from the vantage point of the radical interpreter: that anything we – kindred post-Roman Europeans – can properly be said to *know* about French would have to be accessible, at least in principle, to a Martian radical translator of French – always provided the Martian was a good enough linguist and had enough time to gather the relevant observations. For Quine, any presuppositions we make when translating familiar languages into other familiar languages count as items of known fact – in contrast to, say, convention – only if they could in principle be justified by the methods of radical translation. So if radical translation is indeterminate, then all translation is indeterminate in the sense that the choice between alternative schemes of translation may be beyond justification by appeal to anything factual – anything which may be, properly speaking, known.

As remarked, Quine's view is that it follows from the indeterminacy of translation that meaning itself is indeterminate: if there are no facts of the matter about how an expression may be correctly translated into another language, then there are no facts of the matter about what it means. That may seem a natural enough transition. But it depends, of course, upon an additional assumption: that there can be no facts about meaning which are not accessible to a radical interpreter. And that would seem to involve presupposition of two further, potentially contentious theses:

(1) That there is no first-/third-person asymmetry in the epistemology of understanding: that I can know nothing about what I mean by some particular expression unless you can know it too, by sufficient observation of my linguistic behavior – although it seems clear that, in the typical case where I do know what I mean by an expression, I do not know it by observing my own behavior.[4]

(2) That whatever 'methodology' reconstructs a child's actual learning of a first language, the harvest of that methodology – the understanding of meanings in which pursuit of it results – cannot be *richer* than that of the methodology of radical interpretation. For if it were, there would be a way of knowing facts about meanings which radical interpretation couldn't emulate.

Point (1) may seem attractive, as being merely a version of the thesis that one's meanings must be, in principle, publicly available to others – the thesis of the essential manifestability of meaning which is widely accepted in contemporary philosophy of language (for further discussion, see Chapter 20, REALISM AND ITS OPPOSITIONS; Chapter 11, MEANING AND PRIVACY, and Chapter 12, TACIT KNOWLEDGE). But point (2) may seem less obviously agreeable: doesn't it simply overlook the consideration that the actual learning of a first language may deploy any number of unlearned dispositions – including 'grammatical' dispositions, if Chomsky is right, and also what we might call 'similarity dispositions,' that is, dispositions to find certain aspects of similarity in presented material salient and others not so – of which the methodology of radical translation, if it is characterized along the above austere lines, will take no account?

Each of these reservations would require extended discussion before a considered verdict could be reached on Quine's passage from indeterminacy of translation to irrealism about

meaning. But for the purposes of what follows, we will assume that points (1) and (2) are each sound, and will not pursue such reservations further.

2 Could One Live with the Indeterminacy of Translation?

Let us (provisionally – there are a number of distinctions to be drawn here, which we shall come to below) formulate the thesis as follows:

> For any expression used in a given language, there are at least two incompatible hypotheses about its meaning which equally well – and *unimprovably* – explain all observable aspects of its use in that language.

Suppose this is accepted, and that we are content to conclude from it that, for any two such unimprovable hypotheses about an expression's meaning, there is no fact of the matter as to which of them is correct. Still, it wouldn't seem to follow that there are no facts about meaning at all. The undecidability of the choice between two unimprovable hypotheses is quite consistent with its being definitely *wrong* to choose any of a large number of incompatible alternative hypotheses; the unimprovable hypotheses may be definitely superior to the rest of the bunch, even if the choice between them is underdetermined. So we need to distinguish between weak and strong versions of the indeterminacy thesis:

> *Weak* versions will contend that *some* questions about the meaning of an expression are indeterminate;
> *Strong* versions of the thesis will contend that *all* questions about the meaning of an expression are indeterminate.

Even if we allow that indeterminacy of translation does indeed entail indeterminacy of meaning, it is clearly a strong thesis that is needed if we are to conclude that there are no facts whatever about meaning.

Now, the strong thesis certainly does seem disconcerting; but we should distinguish between better and worse reasons for finding it so. A bad reason would be the thought that all language becomes simply meaningless if we have to take it that meaning is everywhere indeterminate. That's just a confusion: *meaninglessness* is itself a specific–determinate–semantic condition. If there are no determinate facts about meaning, there are no determinate facts about meaninglessness either. But better reasons for disquiet are not far to seek. First, there is the fact that ordinary psychological states, like belief, desire, fear, and so on, which feature so pervasively in our thought about ourselves and each other, are identified by their content: any belief is the belief that p, for some p; any desire is the desire that q, for some q. If there are no facts about the meanings of linguistic expressions, the question immediately arises, how could there still be determinate facts about the content of such states? How could *psychological content* survive an argument which was generally destructive of *linguistic content*? But if it cannot, then it seems that the whole fabric of ordinary psychological explanation must collapse.

Second, if there are no facts about meaning, how can there be facts about *truth*? Our ordinary thinking ascribes truth and falsity to various things: to declarative sentences, to propositions, and to beliefs. But the latter now come under the general shadow which, as

just noted, Quine's thesis casts over the intentional-psychological; and one who regards meaning as strongly indeterminate is hardly likely to be well disposed towards propositions – that is, *reified linguistic contents*. So it is declarative sentences, it seems, which will have to be the canonical bearers of truth-values in the Quinean scheme of things. But then there is the obvious difficulty that the truth-value of a sentence functionally depends both on the way the world is and on its *meaning*. If there are no determinate facts about meanings, it would appear to follow that truth-values are indeterminate as well.

As remarked, Quine himself has not shrunk from the scorn of intentional psychology to which his position appears to commit him.[5] But the concern about the availability to him of ordinary notions of truth and falsity is quite another matter. If strong indeterminacy is sustained, the truth-value of a sentence – if the notion remains legitimate at all – will have to be determined by factors independent of meaning as traditionally understood. A program of naturalized semantics might conceivably prove to be at Quine's service here (see Chapter 8, A GUIDE TO NATURALIZING SEMANTICS), though the prospects for such programs seem anything but encouraging. This difficulty for Quine seems never to have been properly addressed.[6]

3 Quine's Arguments for the Indeterminacy Thesis

Quine has presented two main and quite different styles of argument for the indeterminacy thesis. What has come to be known as the Argument from Below tries to illustrate the predicament of the radical translator by presenting actual concrete alternative translations of certain expressions between which, no matter what behavioral data might be accumulated, it will never be possible to choose rationally. The Argument from Above[7] proceeds, by contrast, on a purely theoretical basis, from the thesis (henceforward the Underdetermination Thesis) that all empirical theory construction is in principle underdetermined by all available data. This second form of argument, in which Quine himself invests the greater confidence,[8] will show, if successful, that translation, and thereby – so we are now allowing – meaning, must be indeterminate even if we lack the wit to construct in detail the sort of illustrations of indeterminacy of which the Argument from Below seeks to provide some examples. In this section we will review some of the twists and turns pursued by the development of the Argument from Below.

As emphasized, Quine is content to grant to the radical interpreter the ability to recognize native speakers' assent to and dissent from sentences formulated in their language, and the ability to interact with them at least to the extent of eliciting assent to or dissent from particular such sentences. In consequence, the translator will be able – except in the most recalcitrant case – to arrive at empirically confirmed generalizations about which types of situation provoke assent to or dissent from instances of a particular sentence-type. Now Quine is content, in *Word and Object*, to resurrect a surrogate for the notion of synonymy which, in "Two dogmas of empiricism," he so roundly rejected: two sentences are said to be *stimulus synonymous* just in case assent to and dissent from them is provoked by the same sensory circumstances. The central contention of the Argument from Below is accordingly that, no matter how ingenious, the translator will never be able to decide rationally between stimulus-synonyms just on the basis of observation of the native speakers' linguistic behavior. Yet intuitively, stimulus-synonyms may be very different in meaning; indeed, they may not even coincide in extension.

Suppose, to follow in the tracks of Quine's famous example, that a rabbit hops past and the translator hears the natives say, "gavagai." Subsequent investigation discloses that the natives are generally disposed to assent to "gavagai" when rabbits are visibly present, and to dissent from "gavagai" when there is no sign of rabbits. So the translator tentatively notes down the translation of "gavagai" as "Lo! a rabbit" or "There goes a rabbit," or something of the sort. That may seem to be a well-grounded translation, but there are in fact a variety of alternatives which the translator would seem thereby to have overlooked. There are, that is, a number of concepts besides *rabbit* which, in exhibiting the observed patterns of assent and dissent, the natives could just possibly be exercising. Some of these are, respectively, the concepts of:

undetached rabbit part,
instantaneous temporal stage of a rabbit,
rabbithood (the universal),
rabbit-fusion (the scattered physical aggregate of all rabbits), and
rabbiting (taken as analogous to the feature-placing concepts, *thunder* and *rain*).

And the one-word utterance "gavagai," could correspondingly mean any of:

There is an undetached rabbit part.
There is a temporal stage of a rabbit.
Rabbithood is instantiated over there.
There is a part of the rabbit-fusion.
It's rabbiting.

The point Quine is making is not merely that the finiteness of the translator's observations must leave open alternative interpretations. The contention is not merely that, however much data the interpreter gathers, there will be rival translational hypotheses which are consistent with it. It is stronger; namely, that there are *certain specific* translational hypotheses such that, however much data the interpreter gathers, each will remain in play if any does.

Now it's clear – however peculiarly the various mooted translations of "gavagai" may strike us – that Quine's point is correct so long as the data to be considered concern nothing but the conditions which prompt assent to and dissent from the one-word sentence "gavagai." However, the thought immediately occurs that the situation is bound to change for the better as soon as we consider more complex sentential constructions in which "gavagai" features as a constituent. For example, suppose we are in a position to put the question, "How many gavagai are there over there?" or "Is that the same gavagai that we saw five minutes ago?" Then the correct answers are bound to vary according to whether or not "gavagai" means: *rabbit*, or: *undetached rabbit part*, or: *stage of a rabbit*, respectively. For one rabbit is many undetached rabbit parts; and stages of a rabbit, unlike rabbits themselves, have no temporal duration.

Quine's reply to this (Quine, 1960, pp. 71–72) is that our ability to run these tests will, of course, depend on our having independently translated certain constructions of the native language as meaning "how many" and "is the same … as." And, he contends, it is quite unclear how one might go about settling the translation of such expressions without *first* settling the interpretation of words like "gavagai" – that is, of sortal predicates – and without identifying the natives' numerals.

That gives pause. Isn't he right? Suppose, for instance, the natives use a word, "qua," which we have come to suspect may be used in concatenation with sortal predicates to ask "how many?" questions. How could we test this hypothesis unless we *already* knew the meanings of a range of such predicates with which it might be concatenated to ask such questions, and could tell whether the answers were as would be appropriate if "how many?" questions were indeed what "qua" was enabling us to put? Indeed there is a further, more specific difficulty. Suppose we have indeed somehow correctly identified the numerals in the natives' language, so that we can tell when the natives are telling us that there is one of a certain kind of object, or three, and so on. And suppose we have settled on the translation of the word "qua," as it occurs in constructions like "qua gavagai," as meaning something like "how many?" And imagine that we put the question, "qua gavagai?" when a solitary rabbit is visible, and get an answer which we rightly take to mean "one." Even so, it is too quick to suppose that the translation of "gavagai" as *undetached rabbit part*, is thereby defeated. It is defeated, of course, if "qua" does indeed precisely mean: *how many?* But "qua gavagai" could, consistently with its eliciting the same answers as "how many rabbits?" mean not that but rather: *Of how many rabbits are there undetached parts over there?* More generally, its role could be this: that if F means *undetached G-part*, then "qua F" means: *Of how many Gs are there undetached parts there?* Under that hypothesis, what looked like a crucial experiment ceases to be so.

Quine's contention, in general, is this: if a pair of expressions have the same stimulus-meaning, then even if they intuitively differ in meaning in ways that would impinge differentially on the use of more complex contexts in which they occur, there will always be a *compensating adjustment* to the interpretation of the surrounding context of such a kind that, under the adjustment, the uses once again coincide. More formally: if F and G have the same stimulus-meaning, but differ in intuitive meaning – like "rabbit" and "undetached rabbit part," for instance – in such a way that, with respect to a particular embedding context, "**Φ**…," the patterns of assent to and dissent from "**Φ**F" and "**Φ**G" could be expected to differ, there will always be an adjusted interpretation of "**Φ**…" such that the assent/dissent conditions of "**Φ**G" under the adjustment will coincide with those of "**Φ**F" when unadjusted.[9]

4 Evans's Appraisal of the Argument from Below

This line of thought, as it stands, is arresting, but hardly sufficiently developed to count as cogent. It deserves thinking through in detail, yet there are few attempts to do so in the secondary literature. However, a distinguished exception is provided by Gareth Evans's paper, "Identity and predication" (1975). Evans contends that Quine looks in the wrong place for considerations that might prove the superiority of the translation of "gavagai" as *rabbit*. Quine's consideration of contexts in which "gavagai" might occur embedded is restricted to what he calls the "apparatus of individuation" – constructions involving identity, plurals, and the numerals. He allows that if we somehow fix on a translation of certain of the natives' expressions within this apparatus, then it will be possible to construct contexts which will discriminate, in principle, among the stimulus-synonyms of "rabbit." His point, then, is that the translation of native expressions into elements of the apparatus of individuation presents a problem which is *coordinate with* that of the translation of "gavagai," and that it is in principle impossible to motivate

the identification of certain native devices as expressing plurality, identity, and so on, without first fixing the translation of terms like "gavagai." Evans's counter is that we do not actually need to consider the apparatus of individuation at all; rather it can suffice to consider how predicates may be used *in combination*, and how they behave under negation.

Here is an illustration of the sort of thing Evans has in mind. Suppose we have identified two other words in the native language, "odolby," and "thewi," which we observe to be associated with the following patterns of assent in the native speech community:

> There is assent to the one-word sentence "odolby" just when something bloodstained is visible.
> There is assent to the one-word sentence "thewi" just when something white is visible.

In addition, suppose we have observed the use of a particle, "neg," which seems to act as an operator of negation; that is, we observe:

> There is assent to "neg gavagai" just when no rabbit is salient;
> There is assent to "neg thewi" just when nothing white is salient;

and so on. And now suppose we also observe the following more complex patterns of linguistic behavior:

> Situations which prompt assent to "gavagai" and "thewi" do not always prompt assent to the conjoined construction, "thewi gavagai." The latter is assented to only when a white rabbit is salient. "Thewi" and "gavagai" will, however, be assented to individually when a brown rabbit sits on the snow.
>
> It is similar with the conjoined constructions "odolby gavagai" and "odolby thewi" – they are assented to only when a bloodstained rabbit is salient, or when something is salient which is both white and bloodstained.[10]
>
> Likewise, the natives are disposed to assent to "thewi gavagai" and "odolby gavagai" when two rabbits are in view, one white, the other bloodstained; but they will assent to "odolby thewi gavagai" only when one and the same rabbit is both white and bloodstained.

Evans's thought is that observations of this character would suffice to eliminate some of the stimulus-synonyms of "rabbit" as adequate translations of "gavagai." For instance, if "gavagai" really were just a device for reporting an environmental feature – like "it is raining" – and "thewi" were the same, then it would seem to be impossible to interpret the conjoined construction "thewi gavagai" as anything other than the conjunction of the ingredient claims: it is whiting and it is rabbiting (compare: it's windy and it's raining). And that translation cannot account for the fact that "thewi gavagai" is not assented to unless the "whiting" is restricted to the surface of a *rabbit*.

The translation of "gavagai" as *undetached rabbit part* also seems to be in difficulties under the hypothesized data. What are we to say that "thewi" means in that case? If it just means *white*, then "thewi gavagai" ought to be assented to whenever a rabbit is salient with an undetached white part – say, a white foot. But that isn't what happens. If the one-word sentence "thewi" means *undetached part of a white thing*, on the other hand, then again

"thewi gavagai" ought still to be assented to when the white-footed rabbit presents himself, since his toes are undetached parts of a white thing, namely his foot. So that translation fares no better. We might surmount that difficulty by allowing "thewi" to mean *undetached part of a white rabbit*; but then we'd be at a loss to understand the natives' assent to it as they gaze at a snowy but rabbit-free landscape.

"Undetached rabbit part" and "temporal stage of a rabbit" are, like "rabbit," and unlike the feature-placer, "rabbiting," sortal predicates. By contrast, the other items on Quine's list of stimulus-synonyms for "rabbit," that is, "rabbithood" and "rabbit-fusion," are singular terms, standing respectively for an abstract and for a scattered concrete object. It is these interpretations of "gavagai" which, Evans contends, are put in difficulty by the kind of data which he envisages for the natives' particle of negation. Suppose our observations disclose that the assent conditions of compound sentences vary depending on the position within them of "neg"; for instance

> "neg thewi gavagai" is assented to whenever no white rabbit is in view, including the case when no rabbit of any kind is in view, whereas "thewi neg gavagai" is assented to only in the presence of rabbits of other colors.

These facts are nicely explained if we suppose that "thewi" means white, "gavagai" means rabbit, and "neg" functions as a device of sentential negation when it takes initial position, and as a device of predicate negation when it immediately succeeds a predicate.[11] But how are the data to be accommodated on the assumption that "gavagai" is, for example, a singular term standing for the universal, rabbithood? On that assumption, the assent conditions of "thewi gavagai" suggest that "thewi" is a predicate of universals roughly equivalent in meaning to "has a white instance here." But in that case we seem to have no way of generating a sentence with the assent conditions hypothesized for "thewi neg gavagai." For the negation particle has nothing smaller to operate on, so to speak, than an atomic predicate of the natives' language; and when it is so restricted, to suppose that it occurs as a predicate negation in that sentence is to predict, falsely, that the sentence should have the assent conditions of "rabbithood does not have a white instance here" – something which should be assented to when there is no rabbit to be seen.

An analogous problem would presumably confront the translation of "gavagai" as *rabbit-fusion*, could we first but find a workable construal in this case of "thewi" as it occurs in sentences like "thewi gavagai." But in fact, as Evans points out, this translation of "gavagai" also inherits the problems associated with the translation *undetached rabbit part*. If "thewi" is a predicate of concrete but spatio-temporally scattered entities, what hypothesis about its meaning will get the assent conditions of "thewi gavagai" right? Not "… has a white part here" – because brown rabbits with white tails don't provoke assent to it – nor even "… has a white, rabbit-shaped part" – because that would leave us bereft of any explanation of the natives' assent to "thewi todagai" in the presence of an Arctic fox.

So far so good, it may seem. Different considerations come into play when Evans comes to the proposed translation *temporal stage of a rabbit*. The sort of difficulty he finds for this proposal (see Evans, 1975, pp. 360–361) is not posed by envisaged data of the foregoing kinds, but has to do with the interpretation of what, in our preferred translation scheme involving rabbits, whiteness, and the rest – henceforward the *favored scheme* – we will naturally take to be simple *tensed* assertions. Suppose, for instance, that the suffix "-p" is naturally taken, in that scheme, as an indicator that a predication is past-tensed.[12]

The question Evans raises is how such data is to be accommodated by a translation scheme for the natives' language which treats the predicates in question as predicates of temporal stages.

Evans's (rather terse) discussion here is semi-technical. He envisages an interpreter who is working within something like the framework of a Tarski–Davidson recursive theory of meaning (see Chapter 2, MEANING AND TRUTH-CONDITIONS: FROM FREGE'S GRAND DESIGN TO DAVIDSON'S) and who first lays down basic clauses which stipulate satisfaction conditions for *tenseless counterparts* of the natives' predicates; for instance

⟨x, t⟩ satisfy "odolby" (tenseless)[13] ↔ (∃y) (y is bloodstained at t & x is a stage of y)

– a pair consisting of a temporal stage, x, and a time, t, satisfy "odolby" if and only if that stage is a stage of something which is bloodstained at that time, and then stipulates satisfaction conditions for stages and tensed versions of those predicates in terms of these basic clauses; for instance, for the simple present tense (where t_u is the envisaged time of utterance)

x satisfies "odolby" (present tensed) ↔ ⟨x, t_u⟩ satisfy "odolby" (tenseless)

– a stage satisfies (present tensed) "odolby" only if the pair consisting of that stage and the time of utterance satisfies (tenseless) "odolby"; and for the simple past tense

x satisfies "odolbyp" ↔ (∃t′) (Before t_u, t′ & ⟨x, t′⟩ satisfy "odolby" (tenseless))

– a stage satisfies "odolbyp" if and only if there is a time earlier than the envisaged time of utterance such that the pair consisting of that stage and that time satisfies (tenseless) "odolby."

Now, an evident effect of the proposed base clause is that *every* stage, x, in the life of a rabbit which is bloodstained at t will be such that ⟨x, t⟩ will satisfy (tenseless) "odolby." So any treatment of the tenses along these lines will have the consequence, as Evans observes, that if any temporal stage of a given rabbit satisfies "odolbyp," or (present-tensed) "odolby," then *every* temporal stage of the same rabbit, *no matter when occurring*, will satisfy "odolbyp," or "odolby" – and indeed will satisfy any other tense of the same predicate, if introduced via a clause along the same lines. Evans evidently regards this kind of promiscuity as a decisive difficulty, for he immediately moves to consider a proposal fashioned to avoid it. But he does not say why. We will return to the matter.

The second proposal Evans considers avoids the promiscuity by an obvious modification in the form of the base clauses; thus

x satisfies "odolby" (tenseless) ↔ (∃y)(∃t)(y is bloodstained at t & x is a stage of y & x occurs at t)

– a stage satisfies (tenseless) "odolby" if and only if it is a stage of something *which is bloodstained at the time at which that stage occurs*. Clauses for the simple present and past tenses may then proceed:

x satisfies "odolby" (present tensed) ↔ x satisfies "odolby" (tenseless) & x occurs at t_u

– a stage satisfies (present-tensed) "odolby" if it occurs at the time of utterance and is a stage of something that is bloodstained at the time that stage occurs; and

x satisfies "odolbyp" ↔ (∃z)(∃t′)((Before t_u, t′) & (z occurs at t′) & (x occurs later than t′) & (x and z are stages of the same thing) & (z satisfies "odolby" (tenseless)))

– a stage satisfies "odolbyp" if it is a later stage of something one of whose earlier stages, occurring before the time of utterance, is a stage of something bloodstained at the time it occurs.

However, Evans foresees a new difficulty for this scheme. According to the proposed clauses, "odolbyp gavagai" should be true just of stages of a rabbit occurring later than a stage in its life when it was bloodstained. What if there no longer *are* any such stages? What if, rather than clean up the bloodstained rabbit, we had destroyed it? In that case, the proposed clause will predict that the natives will no longer assent to "odolbyp gavagai," for there are now no stages to meet the specified condition. But it is easy to imagine how, consistently with the other data envisaged, they might nevertheless give their assent. We have only to imagine that their assent conditions for "odolbyp gavagai" coincide with those of the English sentence, "a rabbit was bloodstained." Evidence of that coincidence, it might seem, would then be powerful evidence for the favored scheme as against the temporal stage-scheme.

5 Are Evans's Objections Compelling?

Seemingly the least cogent part of Evans's discussion is his treatment, just reviewed, of the temporal stage-scheme. Even if the detail of his objections were wholly convincing, there would have to be a vague worry whether the problems thereby disclosed for the stage-theorist were not artefacts of avoidable features of the mooted Tarski–Davidson style of semantic-theoretical treatment of tense. But the detail does not seem convincing in any case. The last consideration, that the proposed clauses cannot recover the assertibility of "odolbyp gavagai" in circumstances when no stage of the rabbit in question post-dates its (last) bloodstained stage,[14] seems crucially to overlook an ambiguity in the hypothesized English (stimulus-)equivalent, "a rabbit was bloodstained." The English sentence can, indeed, be read as embedding a past-tensed predication – when it is taken as the existential generalization, for example, of "that rabbit over there was bloodstained," so that the tensing is done, as it were, within the scope of the quantifier. But the reading germane to Evans's possibility, when "a rabbit was bloodstained" is asserted of a now defunct rabbit, reverses the scope-priority: the past tense is now the principal operator in the sentence, and the quantifier occurs within its scope, so that the effect of the claim is rather that *it was the case that*: [a rabbit is bloodstained]. Any semantic treatment of tense which treats the tenses as operators – on tenseless sentences, in the kind of treatment Evans has in mind, but there are other possibilities – has to be open to all the usual possibilities for ambiguity in the scope of such operators. In particular, we have to expect wide and narrow scope possibilities broadly analogous to those presented by negation. The mooted clause for "odolbyp" is a proposal for a *past-tensed predicate*, where the tense operator is given narrow scope. Evans's objection to it, by contrast, is – irrelevantly – that it does not enable us to recover the (apparent) truth-conditions of predications of "odolbyp" in which it is not merely the predicate but the *whole sentence* that falls within the scope of the past tense. The objection is irrelevant because – their

overt form notwithstanding – such sentences should no more be construed as containing the kind of use of "odolbyp" which the proposed clause concerns than "it is not the case that a rabbit is white" should be construed as containing an occurrence of "is not white." The favored scheme, too, will have to cope with this kind of ambiguity; however it does so, there is no reason to think that the stage-theorist will not have exactly analogous resources.

In any case, what exactly was the problem that moved Evans to dismiss the first form of proposal? Why should the kind of promiscuity imposed by the originally mooted clauses be held to be objectionable? The point might seem obvious. Consider the blood-soaked rabbit of note 12 who gets a thorough cleaning. Suppose the natives assent to "odolby gavagai" before the washing, *but not afterwards*. How, if they are talking about temporal stages of the animal, and if "odolby" has the satisfaction conditions outlined, is this to be explained? If any temporal stage of the rabbit satisfies the utterance of "odolby" before the washing, then all do, no matter when they occur or when the question is raised. So why – if they are talking about properties of temporal stages – do the natives, having earlier assented to "odolby gavagai," not do so later? To be sure, the reference of "gavagai" will then be presumably to stages of the rabbit which post-date its bloodstained period. But, on the treatment proposed, that should make no difference. Since the natives cannot plausibly be taken to have forgotten that the rabbit was bloodstained, or not to have noticed, their temporally selective assent patterns – the very things that motivate viewing them as deploying tenses in the first place – are seemingly at odds with the ascription to them of an ontology of stages and the proposed semantic clauses for "odolby."

This train of thought, however, confuses the form of promiscuity actually entailed by the original type of clauses – promiscuity over stages, as it were – with a form of promiscuity over times. The key point is that those clauses have to be read as dealing with the satisfaction conditions of actual or envisaged *token utterances* of the tensed predicates they concern; that is the effect of their reference to a time of utterance, t_u.[15] The generalization they actually entail is that if a particular (actual or envisaged) historic token of (present-tensed) "odolby," or "odolbyp," is satisfied by a particular stage in the life of a rabbit, then it is satisfied by all earlier and later stages in the life of that rabbit. That is not to be confused with anything which generalizes from a particular historic token predicate's satisfaction to the satisfaction of other tokens of the same type uttered at *other times*. Nothing of that kind follows from the clauses in question when properly construed. In particular, they entail nothing about whether if a particular token of "odolby" is satisfied by a present stage of a rabbit, then that or later stages of the same rabbit should be regarded as satisfying *a later tokening* of "odolby." So there is nothing in the clauses to jar with the natives' hypothesized unwillingness to apply "odolby" to what they know to be a formerly, but no longer, bloodstained rabbit.[16]

I am not suggesting that Evans himself was guilty of this confusion, only that one who did fall into it might find a spurious plausibility in Evans's brisk dismissal of that particular approach on behalf of the stage-scheme – and that I am not sure what else he may have had in mind.

In any case, it must be reckoned as doubtful how forceful are Evans's objections to the temporal stage-scheme. The matter would assume some importance as far as the Argument from Below is concerned if Evans had indeed disposed of Quine's other mooted translation schemes, since the temporal stage-scheme would then represent the argument's last chance, at least as far as Quine himself develops it. But at least one recent commentator has questioned quite generally whether Evans's considerations really are successful (Hookway, 1988, chapter 9). A closer review of the matter will turn out to render further discussion of the temporal stage-scheme unnecessary.

Consider again the data envisaged in order to make trouble for the other schemes. For instance, a brown rabbit with a white foot provokes assent to "thewi" and "gavagai" separately but not to the compound "thewi gavagai." Evans challenges Quine to find an interpretation of "thewi," within the framework of an ontology of undetached parts of things, which explains this. Neither "… is white" nor "… is part of a white thing" will do; "… is part of a white rabbit" – or, more generally, "… is part of a white animal" –would explain why the compound sentence doesn't get assented to in the circumstances described; but it would leave us bereft of any explanation why "thewi" gets assented to on its own in the context in question.

There is an obvious counter. Evans is assuming that we have to find *some one general account* of the meaning of "thewi" to account for both simple and compound occurrences of it. But why shouldn't its syntactic/semantic role be *context-sensitive*? It could, for instance, mean *white*, when occurring in one-word sentences, but *undetached part of a white F*, when occurring in immediate concatenation with the word whose correct translation when not so concatenated is *undetached F-part* (and whose role, when conjoined with "thewi," accordingly reduces to that affixing the parameter, F). Context-sensitive variation in meaning, contrasting with simple ambiguity in so far as the meanings in question are variously cognate to each other, is a familiar phenomenon in natural languages; think, for instance, of the expression "fix" as it occurs in "I'll fix lunch," "he fixed the puncture" and "they fixed the race." Why should not phenomena of that general sort be found in the natives' language too? In effect what would be postulated would be a theory according to which "thewi" had a kind of ambiguity, albeit one in which there was a close relation between its diverse meanings, and where precisely which meaning it took would be determined by its syntactic mode of occurrence. But that doesn't seem outlandish really.

It is similar with the data concerning the use of "neg," represented by Evans as scotching the interpretations of "gavagai" as a singular term standing respectively for rabbithood and the rabbit-fusion. The problem was to find an interpretation of "thewi" which, when "neg" can occur both as external and as internal negation, would rationalize a native's assent to "thewi" when a white rabbit is salient, but to "thewi neg gavagai," when, say, brown rabbits are salient, and to neither when no rabbit is salient. The interpretation "… has a white instance here" captures the first datum; but if "neg" is a device of internal negation, it mis-predicts the assent conditions of "thewi neg gavagai." But again, the obvious rejoinder is that we are not constrained to take "neg," occurring as illustrated, as a device of internal negation in the first place. There could be an operator which, in prenex position, functions as sentential negation, but when it occurs as a predicate-suffix, serves not to negate the predicate – to generate its complement – but operates rather *within* that predicate's content. In particular, under the translation scheme which treats "gavagai" as a singular term standing for a universal, the role of "neg" occurring as suffix may be taken as one of generating the complement of the *adverb* – the mode of instantiation – which we interpret predicates like "thewi" and "odolby" as ascribing. So "thewi neg gavagai" is interpreted as saying not that rabbithood doesn't have a white instance here, but that it has a non-white instance here (as it were, *is instantiated non-whitely here*) – precisely what is wanted to save the data described.

How are we to assess the resulting dialectical situation? It is difficult to see one's way clear to the conclusion that *any* pool of data which a sympathizer with Evans might construct, and which would be *prima facie* recalcitrant for Quine's alternatives to the favored scheme, could be handled in the quite simple kinds of way illustrated. But what is surely convincing in advance is that the most that the sympathizer with Evans is going to be able to do is to call attention to possible observational data which wouldn't square with *particular*

proposed interpretations of some of the expressions concerned; and that it must always be possible in principle to handle such data if one is willing to assign a *variety* of syntactic roles, and/or semantic ambiguities, to the expressions in question. Does this reflection suffice to show that Evans embarked on a lost cause?

It does not. Consider this case. Suppose that alternative schemes along Quinean lines can indeed be constructed which can survive any envisageable addition to our pool of linguistic data, but that whereas the Quinean schemes survive by the postulation of ambiguities of various kinds, the favored scheme has, by and large, no need for such recourse. Then the latter would be, in a clear sense, *simpler* than the Quinean alternatives. Now, the point is well taken that simplicity cannot be assumed, without further ado, to be an *alethic* – truth-conducive – virtue in empirical theory generally. There is *prima facie* sense in the idea that of two empirically adequate theories, it might be the more complex that is actually faithful to the reality which each seeks to circumscribe. But the thought that, when it comes to radical interpretation, there is an ulterior psychologico-semantical reality which an empirically adequate translation scheme might somehow misrepresent is, of course, exactly what Quine rejects – exactly what he famously stigmatizes as the myth of the semantic museum (Quine, 1969, p. 27ff.). And with that rejection in place, methodological virtues which are not, in realistically conceived theorizing, straightforwardly alethic can now become so. In such cases, the methodologically best theory ought to be reckoned true just on that account. It is therefore not enough for a defender of Quine to seek to save the alternative schemes by postulations which, though still principled and general, are comparatively expensive in terms of ambiguity and other forms of complication. If a simpler scheme is available, that fact is enough to determine that these alternatives are *untrue*, by the lights of the only notion of truth that, in Quine's own view, can engage the translational enterprise.

It's another question whether the particular moves I envisaged fall foul of this point. The alternative interpretation of "neg" just canvassed, for instance, postulates a syntactic ambiguity only where the favored, internal/external negation distinguishing interpretation *already* does so, albeit a different ambiguity. And the interpretation of "thewi" – *undetached part of a white F* – to which we had the undetached-parts theorist resort in the attempt to accommodate Evans's assent-data for "thewi gavagai" might actually serve well enough to accommodate the natives' assent patterns to the one-word sentence "thewi" as well ("undetached part of a white something").

That, however, brings us up against a second and this time, I think, decisive consideration, at least if we may take it that the basic clauses of our semantic theory are to assign reference and satisfaction conditions in ways which are *presumed to correspond to the conceptual repertoire of speakers of the language in question*. For even if the schemes considered turn out not to enjoin any avoidable degree of complication in comparison with the favored scheme, the fact is that the range of concepts necessary in order to formulate their various clauses in each case includes, but is not included in, the simple range of concepts of observable spatio-temporal continuants and their observable properties which the favored scheme deploys. So much is obvious in the case of the schemes deploying the concept of the universal rabbithood, and the rabbit-fusion: these are ideas which you do not grasp until you know respectively what qualifies something to be an instance of the universal, and what qualifies it to be a basic part of the fusion. The same point emerges in the clause to which we had the undetached-parts theorist resort for "thewi," and indeed in the various clauses we considered that the temporal-stage theorist might propose.[17]

It is simplicity not in *semantic* theory but in the associated *psychological* theory that is at stake here. Let it be unresolved whether Quine's alternative schemes must issue in semantically more complex theories; it is certain nonetheless that their implied accompaniment must be additional psychological complexity. The effect is that their situation is therefore doubly unhappy. Not merely do they involve the ascription of superfluous conceptual resources to speakers – resources strictly unnecessary to explain their linguistic performance – but, worse, we have to regard the resources in question as lurking behind, but *inexpressible* in, the actual vocabulary of the natives' language. To have the concept of an undetached rabbit part, you need a concept of the integrated individual of which such parts are parts; to have the concept of a temporal stage of a rabbit, you need to grasp the idea of the spatio-temporal continuant of which such a stage is a stage. Yet the Quinean translation schemes will represent you as talking only of undetached parts, or temporal stages; reference to the *integrated, spatio-temporally persisting rabbit* will elude you so long as your expressive resources are fully captured by these translation schemes.

Such schemes, then, even if they can indeed cope with all the data which a sympathizer with Evans might imagine, and even if they can do so without losing out by canons of simplicity governing the construction of *semantic* theory, must, it seems, fall foul of a basic methodological consideration: that the conceptual repertoire which radical interpretation may permissibly ascribe to speakers should exceed what is actually expressible in their language, as so interpreted, only if its ascription to them is necessary in other ways in order to account for their linguistic competence. Perhaps an Argument from Below could be developed in such a way as to respect this constraint. But Quine's own examples do nothing to suggest how.[18]

6 The Argument from Above: Preliminary Clarifications

The Argument from Below operates at the level of sub-sentential expressions. Quine sometimes represents this point by the claim that the conclusion of the argument is not the indeterminacy of translation, properly understood, but rather the *inscrutability of terms*. What a proponent of the argument tries to do in the case of "gavagai," for example, is, as we have seen, to propose hypotheses about its syntactic category and reference in such a way that the truth-conditions – and hence assent-conditions – of contexts containing it are left invariant under compensating readjustments in the interpretation of the other expressions which they contain. Even if this is done successfully, the conclusion will still be consistent, therefore, with determinacy in the matter of what truth-conditions a radical interpreter is to assign to natives' utterances. True, it will be left indeterminate exactly what *thoughts* – individuated more finely than merely by their truth-conditions – should be regarded as expressed by particular native utterances. But the slack will extend no further than the existence of some room for maneuver within assignments whose truth-conditions are the same. It is doubtless for this reason, rather than to acknowledge any infirmity in the Argument from Below, that Quine writes:

> My *gavagai* example has figured too centrally in discussions with the indeterminacy of translation. Readers see the example as the ground of a doctrine, and hope by resolving the example to cast doubt on the doctrine. The real ground of the doctrine is very different, broader and deeper.[19]

Quine thus seems content to have most – perhaps all – of his eggs in the other basket.[20] And the contention of the Argument from Above is indeed stronger. It is precisely that unimprovable translation manuals may differ not merely in their interpretations of sub-sentential expressions, but in the truth-conditions they assign to sentences, and hence in which of the natives' utterances they will enjoin us – in conjunction with our own collateral beliefs about the world – to regard as true.

Here, though, there are a variety of possible theses of differing strength, which we shall do well to distinguish. Let 'M' range over unimprovable translation manuals – by whatever criteria – for the natives' language, and let 'S' denote a particular sentence of that language; let 'C' range over claims which identify the truth-conditions of sentences, and 'c' range over claims which, while falling short of identification, somehow constrain the identification of sentences' truth-conditions, for example, by saying what their truth-conditions are not. Let's say that S's meaning is:

(1) *strongly determinate* if and only if there is some C such that every M makes C about S
(2) *weakly determinate* if and only if there is some c such that every M makes c about S
(3) *weakly indeterminate* if and only if there is no C such that every M makes C about S
(4) *strongly indeterminate* if and only if there is no c such that every M makes c about S

Then any particular sentence S must be in one of three cases:

(A) strongly determinate
(B) weakly determinate and weakly indeterminate
(C) strongly indeterminate

and there are accordingly *seven* possibilities for the sentences, S, of any particular language:

(1) Every S is in case A.
(2) Some S are in case A, some S are in case B, and every S is in one of those two cases.
(3) Some S are in case A, some S are in case B, and some S are in case C.
(4) Some S are in case A, and some S are in case C, and every S is in one of those two cases.
(5) Every S is in case B.
(6) Some S are in case B and some S are in case C, and every S are in one of those cases.
(7) Every S is in case C.

Each of (2) through to (7) represents a possible indeterminacy thesis, of an increasingly radical order.

Now, if a "fact about meaning" is to be anything agreed on by all unimprovable manuals, then it is only (7) which entails that there are no facts about meaning whatever. This is important. I mentioned at the start the underlying physicalist spirit which drives Quine's argument, the perceived difficulty in finding anything for semantical properties to be in what is conceived of an essentially physical world. The thesis of indeterminacy of translation is meant to assuage this concern precisely by showing that semantic facts are *superstition*, and are therefore owed no refuge in the austere ontology of developed physical science. This radical solution will be frustrated should it turn out that Quine's arguments at best support something less thoroughgoing than a thesis of form (7). For in any other situation, there will *be* residual "facts about meaning"; maybe nothing like the rich variety of such facts that

an opponent of Quine would intuitively wish to recognize, but facts about meaning all the same. So we should have the worst of both worlds: insufficient semantic facts to do justice to our intuitions of distinctions of meaning, but enough to set up the perceived difficulty for Quine's physicalism.[21]

The taxonomy invites review of a number of issues. First, on determinacy of truth-value. It might be supposed that, whatever form of indeterminacy thesis is maintained, the effect will be to introduce a species of relativity: the truth-value of a sentence whose meaning is absolutely indeterminate will have to be thought of as likewise non-absolute, as relative to whatever the (unimprovable) manual we happen to favor assigns to it as its truth-condition. But actually there is cause for doubt about the stability of this line of thought. For the truth-value of S can be conceived as determinate relative to some particular manual only if it is thought to be determinate *what that manual has to say* about the meaning of S. And the contents of claims in translation manuals ought, by the Quinean, to be regarded as no more determinate than any others.

If relativism is accordingly eschewed, then it seems it must be conceded that any sentence in category C – any strongly indeterminate sentence – must simply be indeterminate in truth-value, at least so long as we continue to conceive of truth-value as a function of truth-conditions. But how about sentences in category B – weakly determinate but weakly indeterminate sentences? Unimprovable manuals will not converge on the assignment of any particular truth-condition to such a sentence. But they will converge in making certain claims which constrain its meaning, in particular claims which rule out certain such assignments. So the possibility is open, at least while the discussion moves at this level of generality, that such a sentence might yet be determinate in truth-value, since it might be that the space of permissible assignments of truth-conditions is sufficiently narrow to ensure that, as matters happen to stand, the sentence will be true no matter which of those assignments is made. It would all depend how far the weak determinacy extended; how many and what strength of meaning-constraining claims about S the unimprovable manuals would converge upon.

A second issue concerns the respective implications of the various possible strengths of indeterminacy thesis for the viability of ordinary intentional psychology. Assume that it is by the interpretation of what they are prepared to assent to that we are to identify by far the greater proportion of the natives' beliefs. Sentences in category C are obviously useless for this purpose. But might at least some measure of propositional-attitude psychology be feasible if our data concerns acceptances and rejections of sentences in category B? In that event we should know insufficient to determine the truth-conditions of the beliefs thereby evinced. But it is not inconceivable that the range of permissible assignments and truth-conditions on which the best manuals converged might be sufficiently restricted to ensure that, at least in certain special circumstances, the only beliefs which we could regard a particular utterance as expressing would all equally well serve the purpose of rationalizing an associated item of behavior when conjoined with certain plausibly ascribed desires. (For instance, the belief that a certain fruit was nutritious and the belief that it was merely tasty might rationalize many of the same behavioral episodes.) Again, it will all depend on how weak is the weak determinacy involved.

However, I shall not pursue these matters further. The crucial question is: Which of the various possibilities, from (2) to (7), are in the range of Quine's Argument from Above? Here is his own classic statement of the argument:

> Now my point about physical theory is that physical theory is underdetermined even by all … possible observations.… Physical theories can be at odds with each other and yet compatible

> with all possible data even in the broadest sense. In a word they can be logically incompatible and empirically equivalent. This is a point on which I expect wide agreement, if only because the observational criteria of theoretical terms are commonly so flexible and fragmentary. People who agree on this general point need not agree as to how much physical theory is empirically unfixed in this strong sense; some will acknowledge such slack only in the highest and most speculative reaches of physical theory, while others see it as extending even to commonsense traits of macroscopic bodies.
>
> Now let's turn to the radical translation of a radically foreign physicist's theory. As always in radical translation, the starting point is the equating of observation sentences of the two languages by an inductive equating of stimulus meanings. In order afterward to construe the foreigner's theoretical sentences we have to project analytical hypotheses, whose ultimate justification is substantially just that the implied observation sentences match up. But now the same old empirical slack, the old indeterminacy between physical theories, recurs in second intension. Insofar as the truth of physical theory is underdetermined by observables, the translation of the foreigner's physical theory is underdetermined by translation of his observation sentences. If our physical theory can vary though all possible observations be fixed, then our translation of his physical theory can vary though our translations of all possible observation reports on his part be fixed. Our translation of his observation sentences no more fixes our translation of his physical theory than our own possible observations fix our own physical theory.
>
> The indeterminacy of translation is not just an instance of the empirically underdetermined character of physics. The point is not just that linguistics, being a part of behavioral science and hence ultimately of physics, shares the empirically underdetermined character of physics. On the contrary, the indeterminacy of translation is additional. Where physical theories A and B are both compatible with all possible data, we might adopt A for ourselves and still remain free to translate the foreigner either as believing A or as believing B (Quine, 1970, pp. 179–180).

What exactly is the structure of this reasoning? Its premise is clearly indicated. It is the Underdetermination Thesis: the thesis, roughly, that all possible observations – all the observations that scientific observers, however idealized, wherever and whenever situated, might gather between them – do not constrain the selection of an explanatory empirical theory to within uniqueness. Alternative, incompatible theoretical accounts are always possible of any data pool, even if of infinite extent. This is the point on which Quine expects "wide agreement," although he earlier envisages possible disagreement about the level at which Underdetermination operates, some accepting that it holds only for the highest reaches of empirical theory, while others possibly allowing that it go for all empirical theorizing, *tout court*. But how exactly is the transition supposed to be effected to the conclusion: the indeterminacy of translation, and of meaning?

Quine, as the reader will have noted, explicitly disavows that it is simply a matter of applying the Underdetermination Thesis to the special case of empirical linguistics: the indeterminacy is to be "additional." But it is a good question why or whether it would matter much if the argument were indeed that direct. No doubt the direct argument would have to confront a very obvious question: Why is its legitimate conclusion not merely that theories of meaning are, like all empirical theories, *underdetermined* by the behavioral data? Why the additionally strong conclusion concerning *indeterminacy*? Quine shows no inclination to draw the conclusion that empirical theory as a whole is indeterminate. So what, for a proponent of the direct route, would distinguish empirical theories of *meaning*, making the indeterministic conclusion appropriate in their case? It may be doubted that such a philosopher could have any better

answer than to charge that to hold to the opposed view – that meanings may, in cases of indeterminacy of translation, simply lie beyond the reach of empirical detection – is to succumb to the myth of the museum (Quine, 1969, pp. 27ff.). It is to succumb, that is, to the illusion that, in a world apt for complete description by physical theory, there can possibly be states of affairs apt to confer truth and falsity on claims about meaning other than those constituted in behavioral propensities of language use which, by hypothesis, underdetermine the selection of semantic theory. But the salient point is, then, that this – the repudiation of the myth of the museum – is a point on which Quine is going to have to rely *in any case*, even if the argument follows a subtler path than the one disavowed. The immediate conclusion, whatever exactly the configuration of the subtler route to it, is still only going to be that respect for all possible data will leave the translation of the natives' utterances underdetermined. One wonders, then, what exactly the additional subtlety, whatever exactly it may prove to consist in, really has to contribute. Does it somehow make for a stronger conclusion, a more pervasive or deeper kind of indeterminacy? Or does Quine see some difficulty for the simple, direct argument which the subtler route can finesse? What does "additional" mean?

Whatever the answers to those questions, it's notable that the scope of the argument can in any case extend no further than that of the Underdetermination Thesis which fuels it. Someone who allows, for example, that only very high-level physical theory is subject to underdetermination will be under no pressure to concede indeterminacy of translation except for vocabulary which occurs exclusively in such theory.[22] And even for sentences containing such vocabulary, it will be *weak* indeterminacy, not strong, that will be suggested. For however exactly the argument is supposed to run, just as not any old interpretation of that vocabulary would result in a theory which was adequate to the relevant data, so not any old interpretation results in a translation which may justifiably be regarded as reflecting the putatively perfectly rational native scientists' beliefs. The translation of theoretical terms in the native scientists' language can be no more indeterminate than is the selection of an empirically adequate theory of those data.

Strikingly, therefore, Quine's argument promises at best the *mildest* kind of indeterminacy thesis, one of type (2) in the above taxonomy, according to which a thesis of *weak* determinacy/indeterminacy is made out merely for *some* statements. And indeed, even if the Underdetermination Thesis is extended to all empirical theorizing, the most that is in prospect is a thesis of the indeterminacy of theoretical vocabulary relative to some fixed translation of the 'observation sentences.' The argument will have nothing to say about the determinacy of the meanings of the latter; and about the interpretation of theoretical terms, it will suggest only some degree, and by no means an unrestricted one, of latitude.

7 The Argument from Above: Appraisal

Enough of preliminaries. Let us now try to map the course of the purportedly subtler route which Quine officially conceives the argument to follow. It would seem to involve reliance on the following transitional principle:

> If all possible empirical observation underdetermines the choice between theories T1 and T2 (that is, if T1 and T2 are *empirically equivalent*), then a native scientist's responses to his observations will underdetermine the choice between the ascription to him of acceptance of T1 and the ascription to him of acceptance T2.

And that may seem plausible enough. But notice that it does not, by itself, enjoin any conclusions about indeterminacy of translation. It is one thing to suppose that a rational native scientist could quite consistently hold either of two conflicting theories while respecting all possible relevant data. It is another, quite different thing to hold that the sentences by which he expresses whatever theory he does hold may, by an interpreter who respects all relevant data, be translated in different, incompatible ways. The second will follow directly from the first only if the *only* data that the interpreter has to respect concern which data – which observations – the native scientist will have set himself to respect. And that isn't plausible at all, and goes quite unsupported in Quine's presentation. For the project of translation is constrained not just by the need to identify a set of beliefs which, if rational, the native will have arrived at, but to an even greater degree by the need to find plausible *vehicles* of those beliefs in his overt linguistic behavior. Quine's picture of the situation would seem to be that all we – the interpreters – can have to go on in the end is the native scientist's acceptance of certain observation sentences. Quine generously concedes our translation of these, and allows us the assumption that the native is a fully rational theorist of the range of data which they express, so that, to over-simplify rather absurdly, if just two incommensurable but unimprovable theories are possible of these data, then the native is likely to have alighted upon one of these theories in particular. But Quine seems to be depending on the idea that there can be nothing to provide us with *further* guidance in translating the relevant parts of the native's language, nothing additional to motivate viewing it as expressive of that theory rather than of its competitor. And that seems quite unjustified. As theorists of meaning, we will have to locate a *syntax* in those parts of the native's language, and then do a plausible job of mapping the ingredient concepts of one of the theories or the other on to components of his language identified by that syntax, the mapping to culminate in a satisfactory recursive theory of meaning. Quine gives absolutely no reason to discount the thought that the case for one of the interpretations in particular may simply evaporate as soon as this serious work of interpretation gets under way. Bluntly, it may just prove impossible to find the right kind of phonological or morphological structures in the native's theoretical sentences to subserve the necessary lexicography and semantic mapping.

That is one misgiving. We encounter another when we turn to consider just what status the premise – the Underdetermination Thesis – enjoys. Quine wrote that he expected "wide agreement" on this. And, surely, is it not just obvious that theories incorporate more content than the sum of their observational consequences?[23] So isn't it perfectly intuitive that this body of consequences must be theoretically axiomatizable in a variety of inequivalent ways?

When Quine anticipated little resistance to underdetermination, no doubt that was one kind of thought he was having. For instance, let T be some empirical theory and consider two consistent but mutually incompatible supplementations of it, T1 and T2, neither of which entails any empirically testable consequences over and above those of T. Then the choice between T1 and T2 is clearly underdetermined by all possible observations. It merits emphasis, therefore, that this kind of case is *not at all to the purpose*. If Quine's argument is to work then the relevant kind of case has to be one in which, precisely because all possible observations underdetermine the choice between two theories, there is nothing to motivate the ascription to the, by hypothesis, *fully rational* native scientist of one set of theoretical beliefs rather than the other. But equally, of course, if the argument is to work it is essential that the interpreter can have no good reason to suppose that the native scientist accepts

neither theory – essential that there is not a better theory dominating both. And in the envisaged kind of case there will be: for if the native scientist is perfectly rational, he won't be inclined to accept any empirical theory the observational support for which extends no further than for a straightforwardly extricated, otherwise decent enough sub-theory: in the example as envisaged, precisely the theory T.

In brief, gerrymandered examples of Underdetermination, where the incompatibility between empirically equivalent theories is sustained only by their containing empirically idle hypotheses, won't drive Quine's argument. What the argument needs, rather, are cases where empirically equivalent but incompatible theories would either cease to be empirically equivalent, or would lose empirical content, if either was somehow truncated just far enough to eliminate the incompatibility. The clash, in other words, at the theoretical level must be owing to components which are *integral* to the theories' respective capacities to predict and explain the relevant range of observational phenomena. It is not an objection to this point that even when it is required that the axioms be finite in number, any given theory is likely to admit of a variety of axiomatizations, and that difficulties are consequently to be expected for any attempt to characterize precisely which of a theory's components should be reckoned integral to it. Since, as we have just noted, the Argument from Above won't run if the Underdetermination Thesis is made incontestable only by its trivialization, the obligation is actually on the *Quinean* to make out what is involved in the non-trivial case. Whatever it may or may not be possible to say by way of further explanation, what the Quinean requires are examples of pairs of unimprovable theories, the acceptance of each of which in its entirety would be justified on the part of one who knew of sufficiently many of its empirical successes but had no inkling of the other.

With this admittedly vague proviso, what exactly is the Underdetermination Thesis? Again, there are a number of claims of differing strengths to consider. Say that an empirical theory is *tight* just in case it is free of empirical slack of the kind just gestured at, so that it is the underdetermination of tight theories by all possible empirical data that is the material contention for Quine's argument. Let 'S' range over statements whose content potentially befits them to participate in tight theory construction, and let 'T,' 'T*' range over empirically acceptable, global such theories. Then the following are among the possibilities worth singling out:

(1)	$(\forall T)(\forall S)(S \in T \rightarrow (\exists T^*) \sim (S \in T^*))$	*Total theoretical underdetermination*: every component of any acceptable, tight global theory is omitted by another acceptable, tight global theory.
(2)	$(\forall T)(\exists S)(S \in T \,\&\, (\exists T^*) \sim (S \in T^*))$	*Partial underdetermination of any theory*: any acceptable, tight global theory will have some theoretical components which are omitted by another such theory.
(3)	$(\exists S)(\forall T)(S \in T)$	*Partial determination of all theories*: some theoretical statements feature in any acceptable, tight global theory.
(4)	$(\forall T)(\forall S)(S \in T \rightarrow (\forall T^*)(S \in T^*))$	*Total determination of empirical theory*: the theoretical components of any acceptable, tight global theory feature in all such theories.

Now, allowing that the last may be merely utopian, about which (if either) of the first two of these can "wide agreement" be expected? Well, perhaps the history of science throws up some support for an Underdetermination Thesis of type (2). It is possible, for instance, though this is a matter for experts, that Special Relativity Theory and the Lorentzian Theory of Corresponding States share all their testable consequences, and that either might thus in principle be incorporated within an acceptable, tight global theory.[24] If so, then – since no acceptable, tight global theory will contain each of these as sub-theories, but must contain some theory of the phenomena which they explain – thesis (2) may be true. But such local examples seem special at best. It is hard to foresee what argument there might be for something stronger than an Underdetermination Thesis of type (2). Thesis (1), let us be clear, asserts that it is in the nature of empirical theory construction that any tight, empirically adequate, global theory will contain *only* dispensable theoretical claims. What reason is there to think that this is so?

For our present purposes, the crucial reflection is that thesis (3) – that such theories will agree on a common core of theoretical claims – is consistent with thesis (2). So unless the Quinean can make thesis (1) stick, the premise of the Argument from Above, whatever its exact detail, is going to be consistent with the idea that an ideally rational native theorist will be bound, if he is able to take account of sufficiently much of the available data, to arrive at certain *specific* theoretical beliefs, just in virtue of the nature of the project in which he is engaged. And if that were so then, unless there is some special reason to worry about the identifiability of such beliefs, we may equip ourselves, as hypothetically ideal interpreters, with a knowledge of what they are. So equipped, our attempt to interpret the native scientist will be subject to an additional constraint: that of locating expressions for these privileged beliefs among the theoretical sentences which the native scientist is prepared to accept. This constraint may then motivate assumptions about the syntax and meanings of sub-sentential expressions in the native's theoretical language which may rub off on the translation of sentences expressing beliefs of other kinds. In short, it may be an additional source of determinacy of translation.

Those suppositions may, to be sure, be utterly fanciful. The point is only that the Underdetermination Thesis, if it is anything less than the radical thesis (1), is going to be *consistent* with them, and hence cannot validly enjoin any conclusion about indeterminacy of translation in the kind of way Quine seems to have had in mind. Quine's thought, in essentials, was that an assumed knowledge of the meanings of the natives' observation sentences could no more narrowly constrain the interpretation of their theoretical language than the totality of true observations which they could express in that vocabulary would constrain their selection of an empirical theory. We have already had cause for misgivings about the refusal, implicit in this comparison, to acknowledge the routine syntactic constraints to which radical interpretation is subject. But now it appears that Quine has in any case to rely upon what is, so far as I am aware, a quite unsupported and implausible version of the Underdetermination Thesis. To wit, only thesis (1) will do. For once theory construction is allowed to be partially determinate – thesis (3) – ideal interpreters will be constrained to find, among the sentences which a putatively rational native scientist is prepared to accept, some which serve to express the privileged core of empirically determined theoretical beliefs. It cannot be excluded – at least, not without further argument, yet to be provided – that this constraint would greatly reduce their freedom of interpretation, or even that it would have the effect that the interpretational project is uniquely determined. (Maybe each item of the native's theoretical vocabulary occurs in the privileged core.)

Again, I'm not suggesting optimism about such possibilities. It is merely that the premise for the Argument from Above, to the extent that it is something for which support might be forthcoming from the history of science, is consistent with them, and hence insufficient for Quine's notorious conclusion.

* * *

We have found each of Quine's classic arguments, from Above and Below, to provide less than compelling grounds for either the thesis of the indeterminacy of translation or even, more modestly, for that of the inscrutability of terms.[25] Moreover, as stressed at the beginning, Quine's own views have been modified and, in certain respects, softened since he first formulated the arguments on which we have concentrated. Nevertheless, a conviction of the resistability of those original lines of thought is no cause for complacency on the part of friends of the intensional. Although the thesis has been usually received as a paradox, it should be remembered that, within a broader physicalist framework, the indeterminacy of translation would come, at least at first blush, as a relief – the obviation of any need to locate meanings, and intentional states, within a purely physical world. As it is, an abiding tension between the thoughts on the one hand that *in some sense* the world is exhaustively physical and, on the other, that ordinary talk of meanings and the propositional attitudes ought to be unproblematical, remains, and its reconciliation continues to be one of the great issues facing contemporary philosophy.[26]

Notes

1 The reader should be reminded, however, that while this general concern has undoubtedly conditioned and intensified the reaction to the 'skeptical argument' which forms the core of Kripke's interpretation of Wittgenstein, that argument itself – in contrast to Quine's – makes no explicit behaviorist or physicalist assumption.

2 In fact they are directed at different versions of it, as we shall see.

3 For Quine's own original formulations, see Quine (1960, p. 27ff.).

4 Large and subtle issues are raised here. At first blush, it may seem obvious that there are first-/third-person asymmetries of this kind; for instance that, even if my linguistic behavior does underdetermine the translation of my uses of the word "rabbit" – to anticipate Quine's famous example – leaving the radical interpreter with no clearly superior choice among a range of rival interpretations of them, *I* at least can be in no doubt about which, if any, of these rival interpretations is correct. For by "rabbit," I mean of course: *rabbits* – so that's the right interpretation, and anything else is incorrect! But of course the interpreter will expect me to say that. The question, for him, is exactly what knowledge I thereby express. And the question for me – since I would indeed affirm that sentence whatever I meant by "rabbit" – is whether I thereby express any *substantial* piece of knowledge denied to the radical interpreter.

5 Here is a well-known formulation of that scorn:

> One may accept the Brentano thesis [of the irreducibility of intentional idiom] either as showing … the importance of an autonomous science of intention, or as showing the baselessness of intentional idioms and the emptiness of a science of intention. My attitude … is the second…. If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows … no propositional attitudes but only the physical constitution and behavior of organisms. (Quine, 1960, p. 221)

6 Perhaps because it has not been properly appreciated how deep it goes. Someone might think that, so far from posing a problem for Quine, the upshot here – the indeterminacy of truth-value of individual sentences – is merely part and parcel of Quinean holism: the idea, elaborated in §§5 and 6 of "Two dogmas of empiricism," in *From a Logical Point of View* (Quine, 1953), that individual sentences indeed have no meaning except in the context of a larger system – that "the unit of meaning is the whole of empirical science." Indeterminacy of truth-value at the level of sentences may not seem too shocking a matter if it is theories as a whole, rather than their ingredient statements, that are properly conceived as the bearers of truth and falsity.

This suggestion just invites the question, however, of why the whole dialectic does not then replay itself at the level of theories. After all, isn't a theory just a big sentence? So isn't the effect of the holism just to caution against thinking of *small* sentences as the bearers of determinate truth-values? Whereas the problem is to recover, once meaning is indeterminate, *any* space for determinacy of truth-value, even for sentences as big as a global physical theory. If, in company with the indeterminacy of meaning, there is nevertheless to be such a thing as determinate truth-value at any level, then Quine officially needs, for items at that level, an account of truth, and of what determines truth, that liberates the notion from dependence upon any semantic parameter.

7 The Above/Below terminology is Quine's own (1970, p. 183).

8 At least in Quine (1970).

9 The reader may care to think through how the point might apply for **Φ** = "… is the same as…" and F = "rabbit" and G = "temporal stage of a rabbit."

10 Evans seems not to have had a problem with the idea that something might be simultaneously both white and bloodstained! A reader who does will be able to construct another example to make the points about to be illustrated.

11 There will be questions, of course, about its scope where it occurs in the latter mode in sentences involving compound predication; the reader may care to think through what patterns of assent and dissent might motivate particular interpretative proposals about the scope conventions in play in the natives' language.

12 It's straightforward to envisage the sort of data that might prompt the suggestion. Suppose, in full view of a group of native speakers, we take a deeply and thoroughly bloodstained rabbit and wash it completely clean in a stream. Then we put to them each of the following sentences for assent or dissent:

odolby neg gavagai	odolby gavagai
odolbyp gavagai	odolbyp neg gavagai

Finding that the natives assent to both sentences on the left, and dissent from each on the right, would confirm the interpretation of the "-p" suffix as a past-tense indicator. We may suppose this pattern exemplified across a wide range of cases.

13 In order most easily to illustrate Evans's treatment within the framework of the discussion so far, I shall use "odolby" sometimes as present tensed and sometimes as a tenseless counterpart.

14 I am prescinding from the awkwardness, for Evans's purposes, that his objection only engages if the stage-theorist has somehow been stuck with the assumption that the reference of the particular use of "gavagai" is to the last stage in the life of the rabbit in question (i.e., not to an earlier bloodstained stage of that rabbit, or to a stage of a different rabbit). An analogue of that assumption might be more secure if we were concerned with a different example: one whose featured predicate was true only of the last stage in the life of a particular, recently salient rabbit, and of no stage of any other rabbit in the recent experience of our interlocutors. (Evans actually has "running," but that presents the same awkwardness.)

15 This reading is mandated by the reflection that to treat the clauses as concerning *type predicates* instead would rapidly lead to contradictions. Let y be bloodstained at t_1 but not at t_2, and let x be a stage of y. Then $\langle x,t_1 \rangle$ satisfy "odolby" (tenseless), since there is a y such that y is bloodstained at t_1 and x is a stage of y. Let t_u first be t_1. *Then x satisfies "odolby" (present tensed)*. Note that this

upshot involves no relativization to the time of utterance. So now let t_u be t_2. There is no y such that y is bloodstained at t_2 and x is a stage of y. *So x does not satisfy "odolby" (present-tensed).* The italicized claims are overtly contradictory; however, the contradiction is merely apparent if we take it that the two occurrences of "odolby" refer to different tokens of the same type.

Could this contradiction have constituted Evans's objection? No, since it is elicited with respect to a single stage, x, and makes nothing of the generalization across stages on which he remarks. In any case it is easily resolved, as we have just seen, without recourse to anything like his second proposal.

An alternative way to avoid the contradiction for an interpreter who for whatever reason wanted his semantic clauses to concern type-predicates rather than tokens, would be to relativize the notion of satisfaction to times. Such a proposal might proceed with clauses along the following lines (which also avoid recourse to tenseless object-language predicates):

(1) x satisfies "odolby" at time t if and only if something is bloodstained at t of which x is a temporal stage.
(2) x satisfies "odolbyp" at time t if and only if there is some time t′, earlier than t, such that x satisfies "odolby" at t′.

And so on. Note that stage-promiscuity would still be a consequence: such a treatment will entail that if any stage satisfies a predicate at a time, then all stages of the same continuant will satisfy that predicate at that time.

16 Another misgeneralization would confuse stage-promiscuity with a kind of *predicate-promiscuity* – the idea that any stage which satisfies a tensed predicate simultaneously satisfies all other tenses of the same predicate. Predicate-promiscuity would likewise be at odds with the natives' selective use of tensed predicates and, as the reader may care to verify, is indeed entailed, via stage-promiscuity, by Evans's first kind of clauses if they are taken to concern type-predicates. But there is no such implication once those clauses are taken to concern tokens – or are replaced by clauses in which satisfaction is relativized to time (cf. note 14).

17 The development of this point for the case of the feature-placing interpretation is left as an exercise for the reader.

18 This point is also important for the significance of Putnam's permutation argument in his *Reason, Truth and History* (1981, pp. 32ff.); for discussion, see Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM.

19 Quine (1970) opening paragraph. The limitation of the Argument from Below to the inscrutability of terms is expressly recognized at p. 182 of the same paper.

20 Though the reader should note the gist of the concluding remarks to Quine (1970).

21 It is true, of course, that distinctions of meaning which survive Quine's argument will be ones which can be behaviorally grounded and are thus properly public. But that is not enough to make them hygienic from the physicalist point of view. As I stressed at the beginning, the basic worry for physicalism concerning the semantic is its *normativity*, and public meanings are no less normative for being public. A Quinean argument which, while not actually exploiting the presumed normativity of meanings, somehow or other did away with all semantic facts would save the physicalist the task of accommodating this particular province of normativity; but if a residue of semantic facts remains, then so does the problem.

22 Robert Kirk attempts to construct a counter-example to the Argument from Above, exploiting this point: cf. Kirk (1973, p. 198ff.).

23 Strictly, of course, empirical theories issue in categorical claims about observational phenomena only when supplemented with observational premises – statements of "initial conditions." The (intentionally) rhetorical question can be preserved by thinking of the observational consequences of a theory as the corresponding conditional statements, whose antecedents specify the initial conditions, and whose consequents encode the theory's prediction for those circumstances.

24 For detailed discussion of this example, see Zahar (1973, pp. 95–123, 233–262).

25 Always provided, that is, that the issue concerning the latter is not taken to be settled just by the possibility that the assignments of sub-sentential reference effected by an empirically adequate semantic theory for a given language may be varied without loss of empirical adequacy – that is, without loss of consistency with observed patterns of assent. If that possibility is all that is at issue, then the matter is, arguably, settled in Quine's favor by a generalization of the sort of permutation argument offered by Putnam: see Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM. But we have observed that semantic theory has to answer to much more than empirical adequacy in that limited sense.

26 Thanks to Bob Hale, Christopher Hookway, Gabriel Segal, and Jason Stanley for very helpful comments.

References

Works by Quine

1953. "Two dogmas of empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

1960. *Word and Object*. Cambridge, MA: MIT Press.

1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

1970. "On the reasons for indeterminacy of translation." *Journal of Philosophy*, 67(6): 178–183.

1989. "Three indeterminacies." In Barrett and Gibson, 1989, pp. 1–16.

1990. *Pursuit of Truth*. Cambridge, MA: Harvard University Press.

Works by Other Authors

The secondary literature is very extensive. The most useful contributions include the following:

Barrett, R., and R. Gibson, eds. 1989. *Perspectives on Quine*. Oxford: Blackwell.

Evans, G. 1975. "Identity and predication." *Journal of Philosophy*, 72(13): 343–363.

Hookway, C. 1988. *Quine*. Oxford: Polity Press.

Kirk, R. 1973. "Underdetermination of theory and indeterminacy of translation." *Analysis*, 33(6): 195–201.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press. See especially ch. 2, "A problem about reference."

Zahar, E. 1973. "Why did Einstein's programme supersede Lorentz's? I and II" *British Journal for the Philosophy of Science, 24(2): 95–123 and* 24(3): 233–262.

Further Reading

Works by Quine

1974. *The Roots of Reference*. La Salle: Open Court.

1975. "On empirically equivalent systems of the world." *Erkenntnis*, 9(3): 313–328.

1979. "Facts of the matter." In Shahan and Swoyer, 1979, pp. 155–169.

1981. *Theories and Things*. Cambridge, MA: Harvard University Press.

1987. "Indeterminacy of translation again." *Journal of Philosophy*, 84(1): 5–10.

Works by Other Authors

Blackburn, S. 1975. "The identity of propositions." In *Meaning. Reference and Necessity*. Cambridge: Cambridge University Press.

Boorse, C. 1975. "The origins of the indeterminacy thesis." *Journal of Philosophy*, 72(13): 369–387.

Bradley, M. C. 1976: "Quine's arguments for the indeterminacy of translation thesis." *Australasian Journal of Philosophy*, 54(1): 24–49.

Chomsky, N. 1969. "Quine's empirical assumptions." In Davidson and Hintikka, 1969, pp. 53–68.

Davidson, D. 1973. "Radical interpretation." *Dialectica*, 27(3–4): 313–327.

Davidson, D., and J. Hintikka, eds. 1969. *Words and Objections: Essays on the Work of W. V. Quine*. Dordrecht, Netherlands: Reidel.

Follesdal, D. 1973. "Indeterminacy of translation and underdetermination of the theory of nature." *Dialectica*, 27(3–4): 289–301.

Friedman, M. 1975. "Physicalism and the indeterminacy of translation." *Noûs*, 9(4): 353–374.

Gibson, R. 1982. *The Philosophy of W. V. Quine: An Expository Essay*. Tampa: University Press of Florida.

Harman, G. 1969. "An introduction to 'Translation and meaning.'" In Davidson and Hintikka, 1969, ch. 2.

Harman, G. 1979. "Meaning and theory." In Shahan and Swoyer, 1979, pp. 9–20.

Kirk, R. 1986. *Translation Determined*. Oxford: Oxford University Press.

Rorty, R. 1972. "Indeterminacy of translation and of truth." *Synthese*, 23(4): 443–462.

Shahan, R., and C. Swoyer, eds. 1979. *Essays on the Philosophy of W. V. Quine*. Hassocks, UK: Harvester Press.

Stroud, B. 1969. "Conventionalism and indeterminacy of translation." In Davidson and Hintikka, 1969, pp. 82–96.

Finally, *Synthese*, 27 (1974) includes a symposium on indeterminacy and radical interpretation, with contributions from Davidson, Michael Dummett, Harman, David Lewis, and Quine himself.

Postscript

ALEXANDER MILLER

In this brief postscript, I'll argue that Quine's arguments for the indeterminacy of translation survive the objections developed by Crispin Wright in his original (rich and stimulating) entry in the first edition.

1 The Argument from Below I: Simplicity in Semantic Theory

Wright attempts to develop a general argument that will undermine Hookway-style defenses of Quine in the face of the considerations adduced in Evans (1975): that is, defenses of Quine that proceed by assigning "a variety of syntactic roles, and/or semantic ambiguities, to the expression in question" (p. 683).[1] Wright's argument turns on the idea that when empirical theories concern some robustly factual subject-matter, the fact that one empirically adequate theory is simpler than another doesn't entail that it is the simpler of the two theories that captures the facts: on the idea, in other words, that where a robustly factual subject-matter is concerned, simplicity is not a truth-conducive virtue.[2] Given this, and given that Quine's attack on the "myth of the semantic museum" undermines the idea that semantics is robustly factual, the mere complexity of the Hookway-style translation schemes

relative to the simplicity of their Evans-style competitors entails that the latter are true, thus undercutting Quine's argument "from below."

Quine can certainly grant the conditional

(I) If theories concern a robustly factual subject-matter, simplicity is not a truth-conducive virtue.

However, he can question whether Wright is entitled to the desired conclusion

(III) Simplicity is a truth-conducive virtue for theories of meaning.

Attempting to reach (III) on the basis of (I) and

(II) Theories of meaning do not deal with a robustly factual subject-matter

would commit the fallacy of denying the antecedent. In order to reach (III) validly Wright requires

(IV) If theories do not deal with a robustly factual subject-matter, then simplicity is a truth-conducive virtue.

Quine, however, can reject (IV) by arguing as follows. It may well be the case that if a theory deals with some *minimally factual* subject-matter, simplicity can be regarded as a truth-conducive virtue for that theory. However, being minimally factual is only one way a theory can fail to deal with a robustly factual subject-matter: it can also fail to do so by failing to deal with a factual subject-matter *tout court*. Clearly, for such theories, *no* theoretical virtues are truth-conducive.

Wright could perhaps retreat to

(V) If theories deal with a minimally factual subject-matter, simplicity is a truth-conducive virtue

and attempt to reach (III) via the application to semantic discourse of a "disciplined syntacticist" conception of truth-aptitude of the sort he favors: since "'gavagai' means *rabbit*" and the like can be negated, embedded in conditionals, and so on, and are subject to acknowledged standards of correct use, they can be regarded as dealing with a minimally factual subject-matter (Wright, 1992, chs 1 and 2). This suggestion, though, would seem to depend on assumptions about the determinacy of rules (whether syntactic rules or rules concerning "proper use") whose deployment in this context would beg the question against Quine. Arguably, then, Wright's first rejoinder to Hookway's defense of Quine misses its target.[3]

2 The Argument from Below II: Simplicity in Psychological Theory

Wright has another anti-Quinean argument, this time turning on considerations of simplicity in the psychological theory implied by a given theory of meaning:

> [T]he conceptual repertoire which radical interpretation may permissibly ascribe to speakers should exceed what is actually expressible in their language only if its ascription to them is necessary in other ways in order to account for their linguistic competence. (p. 684)

Call this "Wright's Principle." The Quinean translation manual that takes the native to mean *undetached rabbit part* by "gavagai" will ascribe to the native grasp of the concept of an undetached rabbit part and thereby grasp of the concept "of the integrated individual of which such parts are parts" (p. 684). In other words, since grasp of the concept of an undetached rabbit part implies grasp of a concept (*rabbit*) whose instances are integrated, individual rabbits, the Quinean translation manual involves ascribing grasp of the latter to the native, and – crucially, according to Wright – the native will have no way of expressing this concept linguistically given the resources at his disposal. Wright's Principle would thus be violated.

Clearly, there would be no clash with Wright's Principle if the ascription of the concept *rabbit* was in fact necessary to explain some aspect of the native's linguistic competence. So what if the Quinean suggests that grasp of the concept *rabbit* is necessary to account for the native's competence with "gavagai" *as interpreted by the Quinean translation manual*? Wright would presumably reply that what is required is some *independent* justification for ascribing grasp of the concept *rabbit*: that is, some aspect of the native's competence that does not itself presuppose the correctness of the Quinean translation scheme. Equivalently, citing competence with "gavagai" as interpreted in the Quinean manual as the relevant aspect to be explained would be disallowed on the grounds that it is *ad hoc*.

However, disallowing this suggestion on the grounds that it is *ad hoc* is of a piece with the idea that simplicity is a truth-conducive virtue for theories of meaning: the envisaged maneuver is ruled out for introducing complexity into the characterization of the native's competence that is not warranted by some other aspect of his linguistic behavior. However, as argued in §1 of this postscript, it is not clear that Wright is non-question-beggingly entitled in this context to the idea that simplicity is a truth-conducive virtue for theories of meaning.

In any event, even putting that line of argument to one side, it can be argued that the native can in fact express his grasp of the concept *rabbit* via existing linguistic resources. To be sure, "gavagai" on its own won't turn the trick, but "gavagai" *in combination with other expressions in the native's linguistic armory* may well provide the necessary means: for example, "an integrated individual of the kind that gavagai are undetached parts of" refers to integrated, individual instances of the concept *rabbit*. Squaring the Quinean interpretation of "gavagai" with suitable interpretations of the other constituents of this expression in a way that reflects the relevant facts about stimulus meaning may turn out to be non-trivial, but Wright himself (p. 682–683) counsels against the assumption that there will always be some point at which the Quinean will be unable to make the necessary kind of reinterpretative adjustment.

Arguably, then, neither of Wright's arguments – at the semantic level or the psychological level – convincingly block Quine's argument from below.

3 The Argument from Above I: The Underdetermination Thesis

Wright raises a worry about the formulation of one of Quine's key premises, namely, that physical theory is underdetermined by all possible observational evidence. Quine writes:

> Theory can still vary though all possible observations be fixed. Physical theories can be at odds with each other and yet compatible with all possible data even in the broadest sense. In a word they can be logically incompatible and empirically equivalent (1970, p. 179).

Wright argues that the Underdetermination Thesis requires a tighter formulation than Quine provides for, and then that this tighter formulation leaves scope for constraints, not envisaged by Quine, the satisfaction of which narrows down the scope for indeterminacy of translation. In this section I'll discuss the first of these arguments, before turning to the second argument in §4.

Suppose that a fully rational scientist accepts an empirical theory T and that T1 and T2 are two mutually incompatible extensions of T containing no empirically testable consequences beyond those implied by T. Wright concedes that the choice between T1 and T2 is underdetermined by all possible observations, but argues "that this kind of case is not at all to the purpose" (p. 689). According to Quine, a fully rational scientist can choose T1 over T2, or choose T2 over T1, but Wright argues that he must also rule out the possibility that the fully rational scientist accepts *neither* T1 nor T2. And in the relevant kind of case, Wright argues, Quine will not be in a position to do this: the fully rational scientist "won't be inclined to accept any empirical theory the observational support for which extends no further than for a straightforwardly extricated, otherwise decent enough sub-theory: [for example], precisely the theory T" (p. 690).

In response to this, we can argue that choosing T1 or T2 over T needn't mark a departure from full rationality on the part of the native scientist. Suppose that the derivation of observational consequences from T1 utilizes theoretical resources that could be dispensed with at no empirical cost, allowing the derivation of the same consequences from the leaner sub-theory T in a quicker, more efficient manner. Would a fully rational scientist necessarily be averse to T1? It's not clear that he would be. To be sure, if his time is at a premium then it would be irrational for him to rely on T1 rather than T. However, suppose he has time to spare – perhaps even looking for a way to pass the time while waiting for a friend at the café. Given this end, it wouldn't be irrational for him to rely on T1: if he relies on T he may find himself bored and uncomfortable prior to the arrival of his friend. Wright's objection, then, appears to depend, not only on the assumption that fully rational agents necessarily have certain ends (which would be controversial enough in itself), but also on the assumption that those ends would necessarily be frustrated by the choice of an empirical theory whose observational support extends no further than that for an otherwise acceptable sub-theory. This is by no means obvious. In the absence of detailed argument, then, Wright's case against Quine's version of the underdetermination claim appears to falter.[4]

4 The Argument from Above II: Tightness and Indeterminacy

Suppose, though, that Wright's objection to Quine's Underdetermination Thesis stands. Wright suggests that what Quine needs is a version of the Underdetermination Thesis couched in terms of "tight" empirical theories, where "tightness" is characterized as follows. Where T is an empirical theory, incompatible theories T1 and T2 are "tight" extensions of T if they both have observational consequences not implied by T and are both such that if they are "slimmed down" to the point at which they are no longer incompatible they cease to have observational consequences not implied by T. (This allows Quine to avoid the worry discussed in the previous section, since a fully rational scientist who knows T can choose T1 but not T2 or T2 but not T1 without violating any strictures on rationality, but he won't opt to choose neither T1 nor T2 because of the consequent loss of predictive power.)

Can a plausible version of the Underdetermination Thesis be formulated in terms of this notion? Wright suggests that the best Quine can hope for here is: "any acceptable, tight global theory will have some theoretical components which are omitted by another such theory" (p. 690). He then argues (i) that underdetermination – so construed – is consistent with the idea that "some theoretical statements feature in any acceptable, tight global theory" (p. 690) and (ii) that the assumptions about the syntax and meanings of the relevant sub-sentential expressions that we need to make in order to accommodate the core theoretical statements in (i) may, for all that Quine has said, induce determinacy in the translation of theoretical statements outside the core. Since the Underdetermination Thesis – so construed – does not preclude either (i) or (ii) it "cannot validly enjoin any conclusion about indeterminacy of translation in the kind of way Quine seems to have had in mind" (p. 691).

I'll argue that Wright's objection fails to damage Quine's argument, even if we concede all of his claims about what would constitute a plausible formulation of the Underdetermination Thesis.

Quine's argument contains three distinct premises. The first of these is an expression of Quine's physicalism as applied in the cases of translation and meaning:

> PHYSICALISM (PHYS): The facts about correct translation, if there are any, are constituted by the facts about stimulus meaning.

Then comes

> UNDERDETERMINATION (UND): The truth of a physical theory is undeterdetermined by observables in the sense that any acceptable, tight global theory will have theoretical components which are omitted by another such theory.

Next, the principle that translation preserves degree of empirical slack:

> TRANSLATION (TRANS): Insofar as the truth of a physical theory is underdetermined by observables, the translation of the foreigner's physical theory is underdetermined by translation of his observation sentences.

Now, UND and TRANS imply that the translation of a foreign scientist's observation sentences underdetermines the choice of a translation of this physical theory at the relevant level (i.e., the level where underdetermination kicks in).

Since – according to Quine – the meaning of an observation sentence is identical to its stimulus meaning[5] – it follows that the sum total of facts about stimulus meaning (i.e., in this context the sum total of facts *tout court*) underdetermines the choice of translation manual. Hence, there is no fact of the matter as to which translation manual captures the meaning of the foreign scientist's theory.

Quine's claim is thus that PHYS, UND, and TRANS imply the relevant version of the indeterminacy thesis (call this IND). Let CORE be the claim that "some theoretical statements feature in any acceptable, tight global theory." Wright's claim, then, is that the argument from above fails because UND is consistent with CORE and CORE is consistent with the determinacy of translation (not-IND).

This, however, fails to damage Quine's argument. Consider the following argument (where P, Q, and so on are schematic letters):

> P, $P \rightarrow Q$, $Q \rightarrow R$, therefore, R.

Is it any sort of objection to the validity of this argument that P → Q is consistent with T and that T is consistent with not-R? Clearly not. What would be damaging would be the claims that P → Q implies T and that T implies not-R. Given both of these, the set of initial premises {P, P → Q, Q → R} would be inconsistent, showing that at least one of them is false. We would then have no sound way of reaching the desired conclusion R on the basis of them. The same goes for Wright's argument against Quine. Even if we grant that UND is consistent with CORE and that CORE is consistent with not-IND, that in itself poses no worry for Quine. What would be a problem for Quine would be an argument to the effect that UND implies CORE and that CORE implies not-IND. If this were the case the argument as stated could be valid only at the expense of having at least one false premise (since the initial set of premises would imply a contradiction): in other words, its unsoundness would have been established. However, Wright himself makes it clear that he is not suggesting that UND implies CORE or that CORE implies not-IND. He says that he is not suggesting optimism about such possibilities, and indeed describes them as "utterly fanciful" (p. 691). So, for all that Wright has shown, Quine's Argument from Above survives intact.[6,7]

Notes

1 Unless otherwise specified, all references in the text are to Wright's "Indeterminacy of Translation" chapter.

2 Wright himself uses "factual" rather than "robustly factual" in this context, but the idea that what constitutes truth may be minimal in one region of discourse but robust in another is a centerpiece of the approach to realism and anti-realism developed in Wright (1992) and elsewhere, and using it here helps (I think) to see why Wright's rejoinder to Quine in this context is unconvincing.

3 A version of the argument of this section appeared in Miller (2006).

4 Wright could reply that the question about rationality we should be concerned with here is the question whether to *believe* T1 rather than T and then argue further that nothing has been done to suggest that it could be more rational to believe T1 in this case. (Thanks to Bob Hale for raising this worry). However, Quine doesn't need the claim that the fully rational scientist *must* believe T1, only that – consistent with his status as fully rational – he *may*. So isn't the onus on Wright to show that it would be *irrational* for him to believe T1 rather than T? What is the argument for this claim?

5 Observation sentences are sentences whose "stimulus meanings may without fear of contradiction be said to do full justice to their meanings"(Quine, 1960, p. 42).

6 Of course, it would be a problem for Quine if PHYS, UND, and TRANS *collectively* were consistent with not-IND. But since Wright's objection focuses on UND alone rather than on the entire set of premises employed by Quine, there appears to be nothing in what he says to substantiate the stronger, potentially more damaging consistency claim. Wright's references to syntax and the "routine syntactic constraints to which radical interpretation is subject"(pp. 689, 691) won't turn the trick, neglecting as they do the fact that empirical theories are already regarded as consisting of *syntactically structured sentences* at the point at which they fall within the purview of UND and TRANS. Wright is left with the ungrounded, question-begging assertion that the premises of Quine's argument are consistent with the determinacy of translation. (This objection to Wright echoes an objection that can be leveled at Robert Kirk's attack on the argument from above in his (1973) and (1986, ch. 6), as well as his most recent discussion of the argument from above in §10 of his (2004). See Miller (2006, pp. 105–106).

7 Thanks to Bob Hale and Daniel Wee for comments.

References

Evans, G. 1975. "Identity and predication." *Journal of Philosophy*, 72(13): 343–363.
Hylton, P. 2007. *Quine (Arguments of the Philosophers)*. New York: Routledge.
Kemp, G. 2006. *Quine: A Guide for the Perplexed*. London: Continuum.
Kirk, R. 1973. "Underdetermination of theory and indeterminacy of translation." *Analysis*, 33(6): 195–201.
Kirk, R. 1986. *Translation Determined*. Oxford: Oxford University Press.
Kirk, R. 2004. "Indeterminacy of translation." In *The Cambridge Companion to Quine*, edited by R. Gibson, pp. 151–180. Cambridge: Cambridge University Press.
Miller, A. 2006. "Meaning scepticism." In *The Blackwell Guide to the Philosophy of Language*, edited by M. Devitt and R. Hanley, pp. 91–113. Oxford: Blackwell.
Orenstein, A. 2002. *W. V. O Quine*. Chesham, UK: Acumen.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. O. 1970. "On the reasons for indeterminacy of translation." *Journal of Philosophy*, 67(6): 178–183.
Soames, S. 1999. "The indeterminacy of translation and the inscrutability of reference." *Canadian Journal of Philosophy*, 29(3): 321–370.
Soames, S. 2003. *Philosophical Analysis in the 20th Century*, vol. 2, *The Age of Meaning*. Princeton: Princeton University Press.
Weir, A. 2006. "Indeterminacy of translation." In *A Handbook of Philosophy of Language*, edited by E. Lepore and B. Smith, pp. 233–249. Oxford: Oxford University Press.
Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Further Reading

Good book-length introductions or surveys of Quine's philosophy include Orenstein (2000), Kemp (2006), and Hylton (2007), all of which contain chapters on the indeterminacy of translation. See also Weir (2006). For an extended critique of Quine's arguments, see Soames (1999; 2003, ch. 10).

27

Putnam's Model-Theoretic Argument against Metaphysical Realism

BOB HALE AND CRISPIN WRIGHT

Metaphysical realism, as Hilary Putnam conceives it, is not a single, monolithic doctrine, but an amalgam of several closely associated philosophical ideas about the relations between language and reality, and between truth and knowledge or justifiable belief. One component on which Putnam places considerable emphasis is that even an ideal theory (a theory that is '*epistemically* ideal *for humans*' – ideal by the lights of the operational criteria by which we assess the merit of theories) may nevertheless be, in reality, false.[1] But commonly, Putnam presents metaphysical realism as involving adherence to three other claims, of which he takes this feature to be a consequence: that "the world consists of a fixed totality of mind-independent objects," that "there is exactly one true description of the way the world is," and that "truth involves some sort of correspondence between words or thought-signs and external things and sets of things."[2]

The so-called model-theoretic argument has played a leading role in the campaign Putnam has waged, in writings since 1976, against this outlook. Our leading questions will be: What is the argument? How is it best conceived as working? *Does* it work? §I takes up the first, and gives our reasons for concentrating, thereafter, on the version of Putnam's argument set forth in his *Reason, Truth and History* (1981). In §II we explain how, in general terms, that argument is best conceived as working. Cursory inspection of Putnam's overall dialectic reveals it to incorporate three sub-arguments, collectively designed to show that the metaphysical realist confronts an insuperable problem over explaining how our words may possess determinate reference. In our next three sections we expound these three sub-arguments in more detail, and offer some critical reflections on them. §III considers Putnam's version of the Permutation Argument, aimed at showing that reference cannot be determined by fixing the truth-conditions of whole sentences. In §IV we then review his argument that reference cannot be fixed by our intentions or anything else 'in the head'; and in §V we review his 'just more theory' argument, designed to show that the

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

metaphysical realist cannot rescue the situation by appeal to causal or other natural connections between our words and the world. Having argued that the last of these arguments fails, we consider in §VI whether Putnam's dialectical purposes might be better served by other, more specific arguments he has advanced elsewhere, aimed at showing that the project of giving a naturalistic account of reference is hopeless. In §VII we consider how the considerations adduced by Putnam might be seen as an argument telling selectively against metaphysical realism; and we conclude, in §VIII, with a brief assessment of how far Putnam's argument, so viewed, may be taken to succeed.

I

There are significant differences between the versions of Putnam's argument in "Models and reality" (1977) and in *Reason, Truth and History* (1981). Both confront the metaphysical realist with the same challenge – to show how words can stand in the determinate referential relations which his world-view demands. But the latter furnishes the more complete case for thinking the metaphysical realist incapable of meeting it. "Models and reality" deploys the Löwenheim–Skolem theorems, and closely related completeness results, to show – if all goes to plan – that no assignment of truth-values (however tightly constrained) to any (however comprehensive) class of whole sentences can suffice to fix the reference of terms and predicates. But there remain, so far as the argument of "Models and reality" goes, various ways a metaphysical realist may respond: for example, that speakers' intentions or other intentional states play an essential role; or that, if the reference of words is to be thought of as determined via their role in complete sentences, it is not those sentences' truth-*values*, but their truth-*conditions*, that matter. Further argument, of precisely the kind attempted in *Reason, Truth and History*, is needed to close off such moves.

In broadest outline, Putnam's thought in *Reason, Truth and History* has the following structure: If the world is to be conceived as consisting of "some fixed totality of mind-independent objects," with truth consisting in "some sort of correspondence relation between words or thought-signs and external things and sets of things" (Putnam, 1981, p. 49), then there must be determinate referential relations between the words and the things. But if so, the metaphysical realist owes an account of how that can be so. Putnam argues, by reviewing three, putatively exhaustive directions in which it might be sought, that there can be no such satisfactory account:

First, '*what goes on in the head*' *cannot determine what we are referring to*. We can imagine, Putnam suggests, a planet – Twin Earth – very much like Earth, populated by creatures very like ourselves, in surroundings very much like our own. There is, however, an interesting difference – the substance that fills Twin Earth streams and rivers, lakes and puddles, and comes out of Twin Earth taps, and so forth, is not H_2O but has a different chemical composition, XYZ. However, XYZ has just the same phenomenological properties as our water – it looks and tastes the same, and so on, and is, indeed, called 'water' on Twin Earth. If a Twin Earth dweller were somehow transported to Earth, she would not be able to tell our water apart from the liquid she encounters in similar circumstances back on Twin Earth. Her watery thoughts and experiences are, subjectively or 'from the inside,' just like ours. In point of *pure* mental states relevant to the use of 'water' – that is, mental states identified neutrally with respect to the existence and character of such external things as might ordinarily get mentioned in their description – Twin Earth dwellers are indistinguishable from us.

Yet when they speak of 'water,' they are referring to XYZ, whereas we are referring to H_2O. So 'what's in the head' – pure mental states – does not determine reference. Reference varies in a way that cannot be explained by appeal to pure mental states. But to appeal to *impure* – world-involving – states would be just circular.[3]

Second, sub-sentential reference cannot be determined by fixing, via 'operational and theoretical constraints,'[4] either the truth-values *or even the truth-conditions* of whole sentences. This stronger conclusion is now obtained using more modest model-theoretic resources than in "Models and reality." Given one scheme of reference which induces, at each possible world, such-and-such truth-values on complete sentences, we can obtain, by permutation, as many rival schemes as you like, which agree with the 'intended' scheme on the truth-values of whole sentences in *each* world, but diverge over assignments to terms and predicates.

Third, it is no use appealing to any further non-intentional – for example, causal – condition as the needed source of referential determinacy. Any such appeal must assume, for example, that it is at least determinate what worldly relation our word 'causes' stands for. Saying that we use 'cats' to speak of just those things that stand in such-and-such causal relations to our use of the word is 'just more theory' – and, as such, just as liable to unwanted interpretations as anything else we may say.

II

There has been a tendency for commentators to interpret this train of thought as leading to a skeptical paradox comparable to that developed by Kripke in Wittgenstein's name (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §2): as Kripke's skeptic argues that there are no facts about meaning, so 'Putnam's Paradox' would have it that there are no facts about reference, all candidates for the constitution of such facts – the truth-conditions of sentences, speakers' intentional states, and causal and other forms of natural relationships between words and the world – failing to deliver the appropriately determinate goods. It is consistent with such an interpretation of Putnam's argument that he should think, as he certainly does, that the paradox admits of resolution, much as Kripke holds that there can be a solution to *his* Wittgensteinian paradox. But then the suggested parallel begins to limp. For one thing, it enjoins, taken strictly, that any solution will leave in place its skeptical conclusion – that there are no facts about reference – just as Kripke's skeptical solution leaves in place his skeptical conclusion, that there are no facts about meaning. If this were the intended form of Putnam's message, we should expect to find him explaining why/how it is that his preferred *internal realism*[5] can accept such indeterminacy with equanimity. But that is not what he does. What we find is, rather, the claim that the internal realist has no trouble *discounting* unintended interpretations of the sort that plague metaphysical realism.[6] Moreover, while it is true that metaphysical realism requires determinacy of reference – since without it, there appears to be no making sense of the claim that an ideal theory may yet be false – it appears that an outright demonstration of indeterminacy could not tell *selectively* against metaphysical realism. For while internal realism stops short of claiming that even an ideal theory may be false, it will surely grant that a *less than ideal*, but still consistent, theory may be so. And this seems to require setting aside unintended interpretations just as much as does the metaphysical realist's more ambitious claim.

So what is the intended structure of the argument? It might be supposed that Putnam's purpose is not to explode the notion of reference altogether, but by engineering a *conditional*

explosion – by showing that some distinctively metaphysical-realist assumption subserves a proof of indeterminacy – selectively to dispossess the metaphysical realist of the notion. The fact that, as will emerge, no specifically metaphysical-realist assumption oils the wheels of any of the three sub-arguments tells against this line: How, if so, could their combination spell trouble for metaphysical realism but leave internal realism unscathed?[7]

No: the right way to receive Putnam's argument, or so we suggest, is as turning on the crucial claim that *the metaphysical realist distinctively owes an explanatory/constitutive account of reference, but cannot deliver*. Much of what Putnam writes in *Reason, Truth and History* seems to confirm that this is indeed the primarily intended line of attack. The problem about reference, to which chapter 2 in the book is devoted, is repeatedly described as the problem of accounting for *how* the reference of our terms is fixed.[8] The emphasis throughout is on the need for explanation and the metaphysical realist's inability to supply one. Putnam writes:

> Of course the externalist agrees that the extension of 'rabbit' is the set of rabbits … But he does not regard such statements as telling us what reference *is*. For him finding out what reference *is*, i.e., what the *nature* of the 'correspondence' between words and things is, is a pressing problem … For me there is little to say about what reference is within a conceptual scheme other than these tautologies.[9]

The prevailing thought, then, would seem to be that the metaphysical realist incurs certain explanatory obligations which, for the internal realist, *simply do not arise* – that the internal realist may reasonably stay silent when questions are put about the *constitution* of the reference relation, about what makes it the case that a particular expression has the reference it does.

Why this should be so is a matter to which we shall return. But this, we shall assume, is how the overall gist of the argument should be interpreted.

III

Deferring issues about overall strategy, we now, in this section, review some of the detail of, and air some qualms concerning, perhaps the most arresting of the three ingredient claims in the *Reason, Truth and History* argument: the claim that even the *truth-conditions* of whole sentences containing them are insufficient to determine the references of subsentential expressions.

The well-known reasoning in support of this claim affirms that given any domain of objects, and a language used to speak about them, the references/extensions of the subsentential expressions of that language may be permuted consistently with invariance in the truth-value assigned at each possible world to – hence in the truth-condition of – each sentence in the language. This is, Putnam suggests – though the claim needs discussion[10] – a generalization of Quine's contention in *Word and Object* (1960) that reference is inscrutable, based on the so-called 'Argument from Below' (see §3 of Chapter 26, INDETERMINACY OF TRANSLATION). However, whereas Quine merely made a suggestive case that, for all our use of whole sentences containing it dictates to the contrary, 'rabbit' might refer to undetached rabbit parts, or temporal stages of rabbits, or the universal rabbithood – thus posing, at most, an unanswered challenge – Putnam's argument is wholly general ('rabbit'

could, without change in the truth-conditions of any sentence containing it, refer to anything whatever) and, if correct, conclusive.

Let us review the illustration Putnam himself gives of the kind of thing that could be involved in such a systematic reinterpretation. Divide all possible worlds into just three kinds:

(a) worlds in which some cat is on some mat and some cherry is on some tree ('is on' is here tenseless)
(b) worlds in which some cat is on some mat, but no cherry is on any tree
(c) all other worlds.

Now fix the reference of 'cat' in a way which depends on which of these three groups the actual world belongs to. If the actual world is a type (a) world, then 'cat' is to refer to cherries and 'mat' is to refer to trees. If on the other hand the actual world is a type (b) world, then 'cat' is to refer to cats, and 'mat' is to refer to mats. Finally, if the actual world is a type (c) world, then 'cat' is to refer to cherries, and 'mat' is to refer to quarks.[11]

Now reflect that, if the actual world is as a matter of fact an (a)-world, in which some cherry is on some tree, the sentence 'A cat is on a mat' will be true when the references of 'cat' and 'mat' are so stipulated. It will likewise be true in any (b)-world, since those are worlds in which some cat is on some mat, and, in those worlds, 'cat' and 'mat' have their customary reference. Finally, in (c)-worlds, the sentence will be false, since no cherry is on a quark. But these valuations, note, coincide exactly with those of 'A cat is on a mat' as ordinarily understood, with 'cat' and 'mat' assigned their customary reference. In short: the sentence 'A cat is on a mat' could have exactly the truth-conditions it does even if, for some possible worlds, including the actual world, 'cat' were to refer not to cats, but to cherries – or whatever you like.

Putnam shows[12] that this type of maneuver can be complicated so as to embrace simultaneously all the sentences of an entire language. And if sub-sentential reference may be varied in a systematic way without shift in truth-conditions, then whatever – if anything – determines reference, it cannot be the truth-conditions of whole sentences.

This conclusion is apt to seem deeply counter-intuitive. After all, are not the semantics of sub-sentential expressions exhausted by their contribution to the meanings of sentences containing them? So does not the reference of a term, or common noun, say, *have* somehow to be distinctively reflected in the meanings of sentences in which it occurs? The argument gives pause, to say the least. We shall review four broad lines of reservation about it.

One quite common reaction is that Putnam's argument is somehow self-defeating. For in order to receive it as showing the existence of alternative interpretations of a language under which all its sentences retain their truth-conditions, we need already to be able to *grasp* the distinctions, generated by permutation, between the various interpretations. But if we can do that – if we can grasp and distinguish from one another the divergent interpretations on offer – then why can't we just *stipulate* that one among them in particular is the correct one? And why won't that stipulation be sufficient to render reference fully determinate? If, on the other hand, we can't make the requisite distinctions, then we are in no position to follow the reasoning by which Putnam seeks to persuade us that there is a difficulty.

It would be no answer to this to suggest that assumptions of determinacy of reference feature in Putnam's argument only for the purposes of *reductio as absurdum*. They don't. The claim that sentences' truth-conditions underdetermine sub-sentential reference, if

supported by showing how particular permutation-based reinterpretations leave truth-conditions invariant, must depend, for its cogency, on a *continuing* grasp of the differences between the assignments of reference respectively involved in the various interpretations – a grasp which is to survive the drawing of Putnam's conclusion, and on which the grounds for that conclusion depend. So the thought may continue to seem impressive: if we understand the differences, then we can stipulate which interpretation is intended.

The main thing wrong with this objection is, rather, that it misconceives the point of Putnam's argument. It assumes that the argument, like Quine's, is properly seen as a *skeptical* one, directed against the determinacy of reference. Now of course, if that were its project, then the argument had better not proceed in a way which effectively presupposes determinacy, or employs materials which can be straightforwardly exploited so as to ensure it. But that is not Putnam's aim. Putnam's aim is to show, rather, that accounting for the determinacy of reference is a problem specifically for a particular kind of philosophical view. Accordingly – just so long as his own position is not vulnerable to the same difficulty – there is no reason why he should not argue in a way that presupposes determinacy. The intended gist of the permutation argument is merely that whatever secures a determinate reference for a particular sub-sentential expression, it is not the truth-conditions of the sentences in which it features. Putnam's position has to be that if any assumption of determinacy of reference is needed by the argument, it is an assumption which will eventually prove quite innocent from an internal-realist point of view. Of course, it's a good question why or whether that is so, and one to which we shall return.[13]

The second reservation is one some critics misguidedly advanced about the argument of "Models and reality"; that is, that the model-theoretic results to which that essay appeals are applicable only to first-order languages.[14] Strictly, this is so. It suffices to remark, however, that no such concern about generality affects the permutation argument of *Reason, Truth and History*. Given a permutation of the referents of the terms and compensating reinterpretation of the predicates of a language which preserves the truth-conditions – that is, the truth-values assigned at each possible world – of each of its *atomic* statements, it is obvious enough that the truth-conditions not merely of first-order quantifications of those statements but also of their second-order generalizations, and indeed modalizations, will be likewise preserved.[15]

Perhaps more surprising is that the result will also extend to languages containing *intentional* operators, in particular expressions of propositional attitude. One might think that there would be a difficulty here, and that the scope of Putnam's argument would consequently have to be restricted. But this is not so. Take the hardest case: suppose that belief, for instance, is treated as a relation between a thinker and a *proposition*, and that any interpretation is required to assign as referent to a that-clause precisely that proposition which, in view of the assignments that interpretation makes to its sub-sentential parts, the clause in question comes to express. Thus, in Putnam's illustration, 'that a cat is on a mat' comes to refer, in (a)-worlds, to the proposition that a cherry is on a tree and, in (c)-worlds, to the proposition that a cherry is on a quark, whilst keeping its usual reference otherwise (that is, in (b)-worlds). How might we set about gerrymandering an extension for 'X believes that' in order to ensure that 'X believes that a cat is on a mat' retains its actual truth-conditions – is true at just the worlds at which it is actually true – while the referents of 'cat,' 'mat,' and 'that a cat is on a mat' vary in accordance with the permutation in (this extension of) Putnam's illustration?

What is required, naturally, is that 'X believes that a cat is on a mat' should express a truth in all and only worlds in which X believes that a cat is on a mat. Now, since both cases are

logically possible, there will be some (a)-worlds in which X does believe that a cat is on a mat and some in which he does not. We require accordingly that X stands, in just the former, in whatever relation the perverse interpretation assigns to 'believes that' to the proposition that a cherry is on a tree, and fails so to stand in just the latter. Clearly, therefore, we cannot leave the interpretation of 'believes that' invariant. For there have to be (a)-worlds in which X does believe that a cat is on a mat but does not believe that a cherry is on a tree, and in such worlds the truth-value of 'X believes that a cat is on a mat' would accordingly change under the permutation. Hence our reinterpretation will have to assign a new relation to 'believes that.' But what relation? Whatever the relation is, it will have to have the feature that *of necessity* a subject stands in it to the proposition that a cherry is on a tree just when he believes that a cat is on a mat. For if this is not a matter of necessity, then again, there will have to be (a)-worlds in which X does believe that a cat is on a mat, but does not stand in the relation in question to the proposition that a cherry is on a tree; and once again the perverse interpretation will get the truth-value of 'X believes that a cat is on a mat' wrong.

One's first thought is that there may simply be no such relation. But *that is not right*. There is, after all, at least the 'Cambridge' relation, in which a subject stands to the proposition that a cherry is on a tree just in case he believes that a cat is on a mat! One might compare this to the relation in which you stand to Mount Rushmore just in case you have seen a photograph of Snowdon. The crucial point is that permutation-based interpretation works *purely extensionally*. From the extensional viewpoint, the latter relation has been fully specified just when what its extension is has been determined; and we have done that. For all and only our readers who have seen a photograph of Snowdon, the extension is the set of pairs, ⟨You, dear such reader; Mount Rushmore⟩. Similarly, the new relation which our perverse interpretation assigns to 'believes' will be one which has, at each (a)-world, an extension including the pair ⟨X; the proposition that a cherry is on a tree⟩ if and only if that (a)-world is one at which X believes that a cat is on a mat. For (b)-worlds – where 'that a cat is on a mat' is assigned as its referent the proposition that a cat is on a mat – no adjustment in the extension of the usual belief-relation is needed (at least, not in respect of X and the proposition that a cat is on a mat). Finally, at each (c)-world – where 'that a cat is on a mat' is assigned as its referent the proposition that a cherry is on a quark – 'believes' will be assigned an extension which includes ⟨X; the proposition that a cherry is on a quark⟩ if and only if that (c)-world is one at which X believes that a cat is on a mat.

This generality in the scope of the permutation argument is very striking. Arguably, however, the main thing one should conclude from it is how little the kind of 'interpretation' here in play has to do with *real* interpretation, as it were – interpretation in any sense which involves the specification of propositional contents which a thinker might conceivably have in mind. This is, in effect, the area of concern of a different line of objection – what we shall call the 'dilute truth-conditions' objection to which we now turn. The objection concerns the ability of the permutation argument, even if this is sound as far as it goes, to deliver a conclusion of the intended significance. Since the goal of the argument, or so it may seem, ought to be to show that the reference of a sub-sentential expression is underdetermined by any features of the *meaning* of whole sentences containing it, Putnam must implicitly take it that he can encapsulate any germane notion of the meaning of a sentence in that of its 'truth-conditions.' To be sure, talk of 'truth-conditions' is, indeed, a standard philosophical idiom for gesturing at sentences' content. But Putnam's argument, the objection claims, works with so dilute a notion of 'truth-conditions' that this connection is subverted. Putnam's notion requires no more of truth-conditional equivalents than coincidence in

their truth-values in all possible worlds – *strict equivalence*, in the sense of C. I. Lewis – and strict equivalence is intuitively quite consistent with manifest differences in semantic structure and content. In particular, their strict equivalence is insufficient to ensure that a pair of sentences make the same contribution to the content of sentences which embed them. Merely to take a pair of strict equivalents which draw on different conceptual resources – say, 'A and B are parallel' and 'Anything perpendicular to A is parallel to something perpendicular to B' – suffices to open up the possibility of someone who knows one but not the other. It follows that the truth-conditions, hence the content, of – to stay with the particular example – 'X knows that A and B are parallel' and 'X knows that anything perpendicular to A is parallel to something perpendicular to B' also differ. If we assume that the content of those two sentences is determined compositionally, there is then no alternative but to view the semantic contributions, and hence the meanings, of 'A and B are parallel' and 'Anything perpendicular to A is parallel to something perpendicular to B' as likewise different. And if there can be more to the meaning – specifically, the semantic contribution to larger, embedding contexts – of a sentence than whatever it shares with its strict equivalents, then the general thought that the references of sub-sentential expressions may be determined by the *meanings* – in that richer sense, whatever it is – of the sentences which feature them is quite passed over by an argument which shows merely that *truth-conditions*, in Putnam's Lewisian sense, don't determine sub-sentential reference.

How may Putnam reply to this? He had better not challenge the inadequacy of strict equivalence to capture certain finer-grained but still intuitive notions of sameness and difference of sentence meaning. Rather, what he ought to query is the stated characterization of the goal of the permutation argument: we should take the goal of the argument, that is, as that of showing, not that the reference of sub-sentential expressions is underdetermined by *any* features of the semantics of whole sentences containing them, but that sub-sentential reference is underdetermined by any whole-sentence semantical features *which can be explained without prior reliance on specific relations of sub-sentential reference*. That the reference of sub-sentential expressions might yet be recoverable from certain finer-grained semantical properties of sentences containing them – finer-grained than can be captured by relations of strict equivalence and non-equivalence – is accordingly, Putnam may charge, in no way inconsistent with his goal. For what goes into the constitution of such finer-grained semantic properties of a sentence will be, broadly, its mode of composition and the semantics – including reference – of its sub-sentential ingredients. Indeed, it is unintelligible how sentences could have such finer-grained semantic features in the first place unless we simply take for granted a gamut of relations of reference between sub-sentential expressions and items in the world. In short: the reply should be that while there may indeed be finer-grained conceptions of sentence meaning than Lewisian strict equivalence, and while the reference of a particular sub-sentential expression may be recoverable from the finer-grained semantics of sentences containing it, this is all back to front from the point of view of answering Putnam's challenge. That challenge is to explain wherein the determinacy of sub-sentential reference is *constituted*. It is therefore irrelevant to appeal to semantical features of whole sentences which themselves depend upon the reference of those sentences' constituents.

Now, there is a possible misgiving about this reply connected with the question, mentioned earlier, of the extent of the analogy between Putnam's argument and Quine's 'Argument from Below.' The points of analogy are that the conclusion of both Putnam's and Quine's arguments may be expressed in the same way, that we can hold fixed the truth-conditions of a

sentence while varying the reference of semantic constituents within it; and in both cases such constancy of truth-conditions may be glossed as consisting in the fact that, no matter how the world actually happens to be, the sentence will retain – after reference-permutation or Quinean reinterpretation, respectively – the same truth-value as that secured for it by the (presumed) actual reference of its semantic constituents. However, Putnam's illustration would also seem to point to a potentially important difference. The kind of reinterpretation illustrated by the cats-and-cherries example sustains continuity in truth-value only because it is required to be sensitive to *what is actually the case*: for instance, 'a cat is on a mat' is true, under the illustrated reinterpretation, in both type (a)- and type (b)-worlds only because what it says is *constrained to vary* as a function of which, if either, of those types the actual world belongs to. By contrast, any of Quine's alternative translation schemes for 'gavagai' (see §3 of Chapter 26, INDETERMINACY OF TRANSLATION) will construe what the sentence says *in a uniform manner*, no matter what the actual world is like. In short, you cannot tell, under Putnam's assignment of reference, what 'a cat is on a mat' says unless you know how relevant matters stand in the world. But no such knowledge is needed to know the impact on 'gavagai' of any particular one of Quine's schemes. What follows is that for Putnam, but not for Quine, an *additional* distance would seem to be opened between preservation of truth-conditions and preservation of content: precisely, 'a cat is on a mat' retains its actual truth-conditions under the illustrated permutation – that is, has the truth-value it would actually have no matter which of the three types of world the actual world belongs to – only because what it says is *made to change* depending on which type of world that is. And while it may be acceptable for the argument to ignore differences in truth-conditions which can only be specified by presupposing differences in – and hence determinacy of – sub-sentential reference, it is still vital that the notion of 'truth-conditions' which it employs be as strong as possible consistently with that limitation. Yet the notion of sameness of truth-conditions at work in the permutation argument would seem to have *even less connection* with sameness of meaning than strict equivalence does.

This development of the dilute truth-conditions objection probably ought to be open to just the same counter as the original objection. Suppose the objection is right that there is a perfectly good sense in which the effect of a Putnamian – in contrast to a Quinean – reinterpretation will be to have the *content* of a sentence vary as a function of what is actually true. The crucial question, however, is whether this variation in content could be appreciated from a standpoint which takes nothing for granted about sub-sentential reference, but is apprised only of *independently appreciable* semantic properties of whole sentences. What can be known about the semantics of a sentence by someone who knows nothing about the reference of its constituents? Could such a subject know more than its Putnamian truth-conditions, in which possible worlds it would be true and in which false? If not, then the claimed disanalogy between Quine and Putnam would not matter; the permutation argument would still be working with the strongest relevant notion of truth-conditions. Now there are, of course, things other than its Putnamian truth-conditions which someone can know about the semantics of a sentence who does not yet know anything about the reference of its constituents. In particular, there are all the things that are allowed to be available as data for a radical interpreter. Thus it is open to someone who does not yet know the reference of the constituents of 'a cat is on a mat' to observe its *use*, and to note in particular what appear to be its conditions of warranted assent. What he will observe if 'cat' refers to cats and 'mat' to mats is that the circumstances which prompt assent will tend to be those in which some cat is on some mat in a fashion salient to the assentor. But then, as may seem

obvious – indeed, the whole point of Putnam's trick – the same pattern will still be expectable if 'cat' and 'mat' are assigned reference as in his illustration. For suppose that is so and you are asked to assent to or dissent from 'a cat is on a mat.' Isn't it still true that you have only to consider whether you have reason to believe that a cat is on a mat? For if you do, then in both cases – when a cherry is on a tree (in which case that is what the sentence will say) and when none is (in which case it will say that a cat is on a mat) – you will have reason to think that 'a cat is on a mat' is true. So, of course, your observable pattern of assent will be the same.

Prima facie, then, the additional dilution would not matter in any case. However, the decisive point is that, as the proofs in the Appendix to this chapter make clear, its apparently additional dilution of the notion of truth-conditions is actually an artefact of a dispensable – and it has to be said, misleading – feature of Putnam's illustration. There is no need for a permutation-based reinterpretation to 'kink' the assignments of reference after the fashion of the cats-and-cherries example. To be clear about this, consider a specific domain of objects, D, and, for simplicity's sake, restrict attention to all possible worlds involving just those objects and no others. Suppose we have a language, L, fitted to talk about the elements of D and to ascribe a given range of simple properties of them. A permutation of such a domain is simply a one-to-one mapping of D onto itself in such a way that no object need be correlated with itself; and the reinterpretation of the terms and (1-place) predicates of L associated with such a permutation does no more than have each term of L refer to the object onto which the permutation takes its actual referent, and have each predicate of L take as its new extension the set whose members are exactly the objects onto which the permutation takes the objects in its actual extension. Clearly, no matter what the actual extension of a predicate may be, the actual referent of a term will be a member of it only if its correlate under the permutation is a member of the set assigned to that predicate under the permutation-based reinterpretation. Although certain complications have to be finessed to take account of variation on the domains associated with different possible worlds, and of more complex predicates, this simple train of thought captures the essence of the permutation argument. And it points directly to a *uniform* reinterpretation of each sentence 'Fa' of L which is guaranteed to preserve its truth-value in any possible world. Where p is the permutation in question, that reinterpretation will read along the lines of 'the p-correlate of a is a member of the set of p-correlates of Fs.'

We conclude that Putnam has the resources to handle the dilute truth-conditions objection. But there is a related and fundamental worry still outstanding. The immediate effect of the permutation argument is that truth-conditions in, as we have seen, a somewhat technical sense underdetermine sub-sentential reference. And this result, we have stressed, is not to Putnam's purpose unless it bears interpretation as showing that all aspects of the use of a sentence that might be observed without presupposition about the reference of its constituents underdetermine what that reference may be. Now Putnam himself repeatedly expresses his finding as being that reference is underdetermined by both observational *and theoretical* considerations (see, e.g., Putnam, 1989, p. 215). That is a very strong claim. It is tantamount to claiming not merely that alternative assignments of sub-sentential reference are *consistent with* all possible uses of a sentence, but that there will be *nothing to choose between them* even when one takes account of all constraints, beyond empirical adequacy, which condition the construction of semantic theory. This has manifestly not been shown. It hardly seems likely, for instance, that, when all theoretical constraints on interpretation have been reckoned with, there still will be nothing to choose between interpreting speakers as

expressing thoughts of the form: object a is F, and interpreting them as expressing thoughts of the form: the *p*-correlate – for some permutation, *p*, of the domain – of a is a member of the set of *p*-correlates of Fs! (For more on relevant such wider interpretational constraints, see Chapter 26, INDETERMINACY OF TRANSLATION, §5.) In order for the permutation argument to succeed in showing that *best* interpretation of the use of whole sentences always has a variety of schemes of sub-sentential reference to select from, we have to be shown how to find *alternative extensionally coincident thoughts* to correspond to such monstrosities about *p*-correlates and sets of *p*-correlates – alternatives which it is as plausible, in the light of all relevant theoretical constraints, to interpret a subject as expressing by 'Fa' as the simple thought that a is F; and we have to be shown how to do this in a systematic way, right across the language. In short, to make good the suggestion that whole-sentence use underdetermines sub-sentential reference, permutation-based reinterpretations have to be shown to be, by *all* relevant constraints, as good – or anyway to facilitate reinterpretations which are as good – as standard interpretations. The results about permutation, by themselves, are powerless to show that is so.[16]

IV

We turn now to Putnam's argument that the intentional states of speakers are insufficient to determine the reference of their words. The argument, as we saw, proceeds by dilemma: if intentional states are conceived as 'pure' – so that, for instance, speakers both on Earth and Twin Earth can express the very same belief by 'Water is wet' – then reference may vary even though intentional states remain the same. If, on the other hand, intentional states are taken to be impure, so that the content of the belief that 'Water is wet' will be a function of the actual environment of its holders, then beliefs are now individuated by the actual references of the terms that occur in their expression, and thus presuppose, rather than constitute, such facts.

One cause for concern is whether the considerations offered in support of the first horn of Putnam's pure/impure dilemma can be made to cohere with what, later in the argument, he will want to say about the insufficiency of the sort of naturalistic conception of reference to which some – Hartry Field is an actual case[17] – may be tempted in response to Putnam's overall argument. To appreciate, after all, how reference may vary across environments in which the pure mental states of subjects remain the same, one has to have some conception of how reference *functionally depends* on environmental factors. But if such a conception is in place, then won't it constitute at least the beginnings of an account, independently of any play with speakers' mental states or the truth-conditions of sentences, of what it is that does determine reference? – precisely the kind of account which, according to the third stage of Putnam's argument, cannot be given. Unquestionably there is a fair interpretative question here. The externalism about content which the first horn of the dilemma employs is a long-term theme in Putnam's writing; yet the metaphysical realist is apparently to be denied access to this element in Putnam's own philosophy in his attempt to respond to Putnam's challenge. However, we are entitled to proceed without pursuing that question by the consideration that this part of Putnam's argument in any case has no need to proceed in terms of the pure/impure dilemma.[18] A much simpler reflection will suffice. In order for speakers' intentional states, of whatever sort, to serve to establish the references of linguistic expressions, it has to be the case that the objects assigned to those expressions as their referents are

already given as *objects of thought*. It is only as thought about – as referred to in thought – that we can fix, or understand, what it is for a particular object to be the referent of a particular symbol. But the constitutive question being put to the metaphysical realist arises no less for thought than for language. The challenge is to give an explanation of what it is for our thoughts to be of certain objects, rather than others, in the first place. The fact is, accordingly, that there never was any real option of the kind which the pure/impure dilemma is supposed to address. Intentional states cannot constitute reference. That our intentional states already have reference is (an aspect of) the problem, not its solution.

V

If the first two stages of Putnam's argument were to succeed, then the situation would be that no satisfactory constitutive account of reference can proceed in terms of facts concerning our intentional states, or facts about the truth-values, or even truth-conditions, of complete sentences or thoughts. To a metaphysical realist who is also a materialist, however, such conclusions would likely be entirely congenial, merely serving to underline the need for a quite different account of reference in broadly naturalistic terms. That it would be quite mistaken to think that any such account could meet metaphysical realism's needs is the burden of the third component of Putnam's argument.

This comprises, in fact, several distinguishable lines of attack: some of them are directed specifically at the idea that reference can be fixed by causal connections, but others aspire to greater generality, purporting to establish that there can be no 'reductive' explanation of reference in naturalistic terms or, more generally still, that once it is allowed that neither intentional states nor truth-conditions can form the basis of an explanation of how reference can be determinate, it can be seen that nothing else can do so either. The concern of this section will be with the most general – and most notorious – such line of all.

Putnam writes:

> Suppose there is a possible naturalistic or physicalistic *definition* of reference, as Field contends. Suppose
>
> (1) *x refers to y* if and only if *x bears R to y*
>
> is true, where *R* is a relation definable in natural science vocabulary without using any semantical notions … If (1) is true and empirically verifiable, then (1) is a sentence which is itself true even on the theory that reference is fixed as far as (and *only* as far as) it is determined by operational *plus* theoretical constraints ….
>
> If reference is only determined by operational and theoretical constraints, however, then the reference of '*x bears R to y*' is itself indeterminate, and so knowing that (1) is true will not help. (Putnam, 1981, pp. 45–46)

Knowing that all instances of (1) are true won't help, Putnam thinks, because, by the permutation argument, they will remain true – and, indeed, will have the same truth-values in all possible worlds – when 'R' is taken instead to stand for a quite different relation R*. In fact, there are as many such alternatives R* as there are permutations of the universe of discourse. So supposing reference to be R has no more explanatory merit than supposing it to be R*. Hence it is merely an illusion that a unique reference relation has been singled out.

This move – of holding any attempted naturalistic characterization of reference to be 'just more theory' hostage to permutative reinterpretation – is one that Putnam repeatedly

makes in the closing stage of his various attacks on metaphysical realism.[19] If allowable, it is of course decisive – for it will be available against *any* specific constraint the metaphysical realist may propose, regardless of its precise content, just so long as the constraint is formulated in a language to which the permutation argument applies.[20] The obvious and crucial issue is: Is the move fair, or foul?

Well: it is foul, for a reason first stressed by David Lewis (1984, pp. 224–225). There is a distinction to be made between, on the one hand, an interpretation's *modeling* a proposed constraint – *making a statement of the constraint come out true* – and on the other, the interpretation's *actually conforming to that constraint*. The 'just more theory' gambit seems simply to miss this crucial distinction, taking the former for the latter.

To elaborate a little: Let C be some proposed (naturalistic) constraint on reference generally, L a language, S a sentence of L expressing C, and I an interpretation of L. Suppose that I does indeed induce the value *true* on S. It might seem that I must conform to C; for S expresses C, after all, so that if I makes S come out true, isn't that just the same thing as I's conforming to C? Well obviously, not at all: we have no right to assume that, *whatever interpretation* of L is in play, S will (still) express C. Suppose, schematically, that S has the form:

$$\forall x \forall y \forall z(\text{speaker}(x) \,\&\, \text{expression}(y) \,\&\, \text{object}(z)) \rightarrow (x \text{ refers to } z \text{ by } y \rightarrow R(x,y,z))$$

One thing that may vary under different interpretations of L is, naturally, the relation assigned to 'R.' We may not take S as expressing C *tout court* – some interpretations will have S expressing C, others won't. An interpretation J under which S fails to express C may still make S come out true. And if we are able to express C in some other language L*, with resources sufficient to discuss the semantics of L, we may be in position to state – and it can be true – that while S is a sentence of L true under J, J does not conform to C.

A supporter of Putnam might reply that this will be a situation we can recognize as obtaining, and to which we can give expression, if, *but only if*, we can fall back on some other language L*, the reference of whose expressions can be assumed to be (sufficiently) determinate – in particular, there will have to be a sentence of L* by means of which we can give determinate expression to the intended constraint C. And it is precisely at this point, it may be alleged, that the metaphysical realist runs into 'just more theory' trouble. For his predicament is that any language, L* no less than L, will raise just the same problem about determinacy of reference. He can't just assume a more inclusive but referentially determinate L* in which it may be asserted that whilst the sentence S does indeed come out true on a whole host of interpretations (of L), all but a few of these are ones under which S fails to express the proposed constraint, to which they, furthermore, fail to conform. And if he can't simply assume that he can convey this thought in words, he cannot assume that he can think it either.[21] The upshot is, the supporter may claim, that while there is indeed a distinction of the sort Lewis proposes, the metaphysical realist cannot avail himself of it in the situation which matters, when any metalanguage, no less than the language with which we are originally concerned, gives rise to just the same difficulty.

But if this is the best reply that can be made, Lewis is right to cry "Foul!" Just consider the dialectical situation. The metaphysical realist – Field or Devitt, for example – takes up the challenge to say what constitutes determinate relations of reference, only to find that no sooner has he opened his mouth than Putnam gags him with the complaint that he has no right to assume any of his words to be determinate in reference. The resulting situation is therefore really no different from that generated by the boring and jejune variety of

meaning-skepticism which challenges an opponent to explain how meaningful discourse is possible, but won't countenance attempted answers because to presume them meaningful is to beg the question against it. Obviously the metaphysical realist has to be presumed capable of contentful – so, determinately referential – speech if he is to respond to Putnam's challenge, or indeed to any challenge at all. The onus legitimately placed upon him is not to *demonstrate that* determinate reference is possible, but to provide a constitutive account which *explains how* determinate reference works. Accordingly, he is perfectly within his rights to assume, at least *pro tem*, a metalanguage in which a determinate account of the putative mechanics can in principle be given.

VI

If the 'just more theory' move is illicit, that need still be no very serious matter for the overall argument, provided there are good independent reasons for doubting that any naturalistic reduction of reference can be provided. Putnam has assembled in different places a variety of more specific arguments to this conclusion, of – so it seems to us – somewhat differing levels of cogency. We shall briefly review two lines in particular.

The first occurs in "Model-theory and the 'factuality' of semantics" (1989). The form of naturalistic proposal Putnam there envisages is familiar from such natural scientific identifications as those of water with H_2O or of heat with mean kinetic energy of molecules. By identifying heat with mean molecular energy of motion, we accomplish what seems to be the best available explanation of empirically attested correlations involving variations in the temperature and pressure of a mass of gas whose volume is kept constant and so forth, and take this to sanction the identification. Might it not be, likewise, that by identifying the relation of reference with a certain physical relation, R, holding between tokenings of expressions and the worldly items to which they refer, we may achieve the best explanation of certain aspects of our use of those expressions? That is, why should ordinary scientific methodology not turn out to provide the same kind of case for identification of reference with R as for the identification of water with H_2O, or heat with mean molecular motion? (cf. Putnam, 1989, pp. 216–217).

Putnam's objection[22] is that any such proposal, grounded upon explanatory virtue, is viciously circular. Here is a key passage:

> One difficulty … is that this [proposal] uses the notion of *truth*. Our problem … was to explain how a particular reference relation – and that means, also, a particular extension for the notion of truth – gets attached to our words. To say that what does the attaching is the fact that certain sentences … are *true*, … is flagrantly circular. The problem, of course, is that what the semantic physicalist is trying to do is reduce intentional notions to physicalist ones, and this program requires that he not employ any intentional notions in the reduction. But *explanation* is a flagrantly intentional notion.[23]

We can discount what may seem to be the principal complaint in this passage. The general shape of the type of proposal mooted is that it is because a certain physicalistically specifiable relation R holds between our words and their referents that those words do in fact have those referents. It is, therefore, simply a misrepresentation to treat the proposal as asserting that the fact that certain *sentences* are *true* is what explains why our words refer as they do.

That is, it seems quite gratuitous to impute to the physicalist the contention that what 'does the attaching' is the fact that a certain sentence (saying that our words bear R to some object) is true, rather than (simply) the fact which that sentence purports to state. This indifference to the distinction between object- and metalinguistic claims merely invites repetition of the main complaint already leveled at the 'just more theory' move.

The point about explanation made in the second half of the passage may seem more telling: if the semantic physicalist is in the business of giving a reductive account of reference in particular, and intentional notions in general, how can it be permissible to deploy intentional notions in so doing? But this too seems to us of doubtful force. Maybe the question would be appropriate if what was at issue was the standard type of *analytic* or *conceptual* reduction, a purported analysis of necessary and sufficient conditions of application. But the mooted form of proposal actually seeks an *a posteriori* reduction. It is anything but clear that all use of the notion of explanation must be eschewed, if the aim is that of saying what naturalistic relation between words and things in fact underpins reference – if what's on offer is a *theoretical* identification of reference with R, of the same general character as the identification of heat with mean molecular motion.[24] All it seems to be legitimate to impose, by way of a general constraint, is rather that if use is made of an intentional notion in a statement which is part of a program of physicalistic reductions of intentional notions generally, that use must be of a kind ultimately amenable to that form of reduction. It would be necessary to look at the details to see whether a particular physicalistic construal of reference, or explanation, violated this rather vague constraint. In any case, the constraints on legitimate *a posteriori* identification of properties and relations still remain to be clearly worked out.

Putnam is on much stronger ground, however, if it may be assumed that the naturalist proposal must ultimately identify reference with some *specific form of causal relationship* between the item or items that stand as the reference of a term and token uses of that term. Putnam himself, of course, has been prominent among those who have emphasized, as against the once orthodox Fregean conception of the matter, the role of causality in the determination of reference in a wide class of cases. But he warns us that neither his own proposal, nor Kripke's similar idea, were intended to explain from a standing start, as it were, how determinate reference is constituted; clearly they could not do so, since both pictures simply assume from the outset that individuals can be "singled out for the purpose of a 'naming ceremony,'" and say nothing about how that might be done *ab initio*.[25]

Putnam has expressed various doubts about the viability of a reductive causal theory of reference. As he stresses, it will normally be the case that very many of the objects and events that figure in the causal ancestry of a particular utterance of an expression will not be what it refers to. Further, a term's or predicate's reference may be to, or may include, things with which it is not causally linked – items existing only in the future, for instance, are presumably available to be referred to but as yet sustain no causal relations.[26] Part of the problem for the causal theorist, then, is to single out the right causal relationship. Putnam is skeptical that this can be done in purely naturalistic terms without falling back on intentional notions. As against Evans's version of a causal theory, for example, according to which, roughly, a term refers to the dominant source of our beliefs involving it, he justly observes that the dominant source of our beliefs about electrons, say, may well be physics textbooks, rather than electrons themselves.[27]

Obviously these considerations are not conclusive. To the difficulty about future things, for example, it may be replied that in cases where the term introduced is general (perhaps a natural-kind term) it is to be understood that its extension comprises the causally connected

samples and all other things of the same kind.[28] In general, causal theorists will surely agree that work is needed to characterize the appropriate kind of causal link – but why suppose the project to be hopeless?

Well, we suspect the project is hopeless. The core difficulty is to restrict, without ineliminable play with antecedent assumptions about its reference, the utterly disorderly mess of items that are apt to elicit tokenings of any given expression. In his Gifford Lectures (Putnam, 1992, ch. 3) Putnam discusses probably the most sophisticated attempt to date to accomplish this: the proposal of Jerry Fodor (1990) that the extension of a term comprises the smallest class of items which as a matter of natural law cause tokenings of the term, and whose doing so asymmetrically explains all other tokenings of the term. For example, both horses and pictures of horses are apt to cause tokenings of 'horse'; but Fodor's intuition is that horses are the basic cause and therefore qualify as the reference, since it is only because horses cause tokenings of 'horse' that pictures of horses do.

Against this, it may be objected that there really is no clear priority as between 'If horses did not cause tokenings of "horse," neither would pictures of horses' and 'If pictures of horses did not cause tokenings of "horse," neither would horses.' Rather, what seems to be true is that it is because 'horse' refers to horses that *both* horses *and* pictures of horses – and thoughts of horses, and cows in a darkened field, and so on and so forth – elicit, *ceteris paribus*, tokenings of 'horse'! In the jargon of possible worlds, the closest worlds in which pictures of horses do not cause tokenings of 'horse' are worlds in which horses don't either.[29] (For a fuller discussion of difficulties with semantic naturalism, see Chapter 8, A GUIDE TO NATURALIZING SEMANTICS.)

VII

Let us try to take stock. First, to summarize the situation of the three sub-arguments of *Reason, Truth and History*. That the metaphysical realist has no option of explicating reference in terms of intentional states we take to be clear. However, the claim of the permutation argument to have shown that reference is underdetermined by features of the use of whole sentences is, as we saw, open to question. Moreover, the 'just more theory' move *is* a foul, and some of Putnam's own specific criticisms of causalist/naturalist proposals about reference are less than conclusive. However, to observe that the permutation argument as it stands is inconclusive for Putnam's purpose is one thing; but to make the kind of positive, constructive case for the determination of reference by whole-sentence semantics which, if such was her strategy, the metaphysical realist would need, is quite another. It is no clearer how such a case might in detail be made. Moreover, if that is not to be the strategy, then a causal account of reference – broadly construed – is the only remaining avenue to explore, yet the literature justifies nothing but pessimism about reconstructibility of semantic notions in non-intentional, causal terms. Putnam, then, may not have strictly proved all of his three lemmas. But he has done enough to issue a very pointed challenge, and one to which it is by no means clear that the metaphysical realist can satisfactorily respond.

Second, it merits emphasis that Putnam's considerations, even if conclusive, would provide no argument for the indeterminacy of reference as such; rather, what they would establish is that *if* referential relations had to be constituted in a certain kind of way – in the truth-conditions of sentences, for instance, or in causal connections – *then* reference would be indeterminate. The proper conclusion would be merely that a constitutive account of

reference, of what makes it the case that a particular term, thought or spoken, stands for a particular object or kind, cannot proceed along any of the three lines reviewed. If those lines indeed exhaust the possibilities, then a case would have been made that there can be no fully explicit, reductive account at all of what constitutes the reference of a symbol to any particular item or range of items – at least, none which does not take for granted the determinacy of reference of our thoughts as a background. 'Aboutness' would have to be conceived as primitive.[30]

Such a finding would no doubt be of great interest. But it will have been achieved, if it can be achieved, in a way that has no evident selective bearing on the status of metaphysical realism. The argument, if it can be made good, will be an argument for everybody. Moreover, notions which promise to admit of no reductive account are anyway ten a penny. So the questions remain: Why see in the situation a (potential) *problem* to do with reference? And why, if so, a problem *distinctively* for *metaphysical realism*? The crucial task for a would-be sympathetic interpreter of Putnam is to provide convincing answers to these questions. How might such answers run?

To lack a constitutive account – and all prospect of a constitutive account – of a certain kind of subject-matter is not, except in special circumstances, to have a reason to distrust its reality. That Putnam himself intends no skepticism about reference is abundantly clear from his willingness to allow that we can perfectly legitimately and fully adequately specify what 'cat,' for instance, refers to: its reference is to *cats* (and therefore not to cherries, or to the *p*-correlates of cats under some permutation, *p*)! More generally, if it is granted that the language in which we are to state the reference of a term is an extension of the language to which that term belongs, then a homophonic formulation is a perfectly adequate response to someone who challenges us to individuate the reference of that term. If, however, that assumption is not granted – that is, if object- and metalanguage are distinct and the challenge is to justify the assignment of one scheme of reference to the terms of the object language, rather than to a permutation of it – then there are perfectly ordinary canons of interpretation to justify a preference, for example, for the assignment of cats to be the extension of 'chat' in French, rather than cats* – that is, cherries in a world in which cats are on mats and cherries are on trees. These will be canons which have to do, for instance, with the salience of cats in many of the situations which provoke 'chat'-talk among the French and a corresponding salient absence, for the most part, of cherries. That there are correct and incorrect things to say about what expressions refer to is enough for there to be *truths* – at least on the conception of truth favored by the internal realist – about reference.

This is the key to the question of the selective bearing of the argument. What, precisely, might be put in doubt by the kinds of consideration reviewed is the existence of truths about reference *in a more substantial sense of* 'truth,' a concept of truth whose applicability to claims of a certain kind requires, beyond the unimpeachability of those claims in the light of the ordinary discipline that informs their use, some form of robust *fit* between them and the world. For it is not enough for metaphysical realism merely that there be facts about what the expressions of our language refer to: these facts must be facts as metaphysical realism is wont to conceive *all* facts, facts no less sublime than – since constituted by relations to – the sublimated objects and properties which make up the metaphysical realist's world. There is accordingly no question of resting content with the sort of deflated account of them which is all that is provided by the homophonic platitudes and routine methodology of interpretation for which the internalist about truth may settle.

The metaphysical realist, then, owes a perspective on the nature of relations of reference which allows them to stand behind the routine interpretative methodology and which, indeed, explains its adequacy – explains how it is indeed a way of 'getting onto' or 'tracking' these independently constituted relations; a perspective which allows us to construe the truth of ascriptions of reference along robust correspondence lines, and which generally finds a place for such relations in the world as metaphysical realism conceives it. And there is, if Putnam's argument can succeed, no such perspective possible, because there is then nothing to be said about what reference *is*.[31]

In brief, then, we have a rich and complex argument to the conclusion that reference admits of no reductive account, coupled with the claim that metaphysical realism – but not internal realism – is saddled with a world-view that cannot be properly understood unless such an account can, *per impossibile*, be given. The crucial difference is entirely one of explanatory obligation. For metaphysical realism, reference is a matter of relations between robustly distinct existences, items of language and thought on the one side, and items in a stubbornly alien world on the other; and this conception, Putnam's driving idea has it, entrains a commitment to the possibility of some sort of external perspective on the nature and constitution of this relationship – exactly what, if his argument succeeds in detail, cannot be delivered. So the metaphysical realist must, in the end, be driven to obscurantism: a conviction in the reality of relations constituted, he knows not how, between his thought and a world wholly alien to it.[32]

VIII

Why does internal realism incur no parallel obligation? Can the mere currency of standards of correctness for claims about reference really ensure that no issue arises? It is one thing to get a sense of Putnam's thought on this point, another to determine whether it is really convincing. The key idea seems to be that, as Putnam repeatedly expresses it, 'there is no ready-made world': that the division of the world into particular objects and kinds of thing is somehow coeval with, rather than merely *reflected by*, the divisions among our concepts and the expressions for them. If the kind picked out by a term of ours is thought of as originally constituted quite independently of the use of that term and the conceptual resources associated with it, then the question has to arise: What attaches the term on to just that kind, as opposed to another? That is the question which metaphysical realism is charged to answer. If, by contrast, the kind is regarded as in some way having no being independently of our deployment of those very conceptual resources, then there is no real linkage to explain, any more than it wants an explanation, how the patterns on a slide manage to be congruent with the images it casts upon a blank screen.

This kind of simile is convincing enough in its way. The difficulty is to give it substance in the case that matters – to see what the idea that human conceptual activity 'slices up' the world really comes to.[33] But perhaps on reflection there is room to repudiate the metaphysically realist conception of a 'hooking' of language onto a sortally predeterminate world without recourse, natural though it may be, to opposing constructivist metaphors.[34] The crucial point is that, unless the unity of a range of items is in some way fixed in advance of the institution of using a term of which they are the reference, there is no non-trivial question what makes for the connection between that term and that range: the range of items in question just constitutionally is that for which the term in question stands.

That leaves the metaphysical realist the options of faulting the detail of the stages of the argument, or living with its conclusion: that to conceive of the world in a certain kind of robustly autonomous fashion is to consign the relation between the vehicles of our thought and the taxonomy of the world to unaccountability. Putnam effectively ridicules such an upshot. But ridicule, it may be countered, is no substitute for argument. Any broad philosophical system will have its primitive notions and theses. Further argument may be demanded as to why metaphysical realism may not legitimately go primitive at the interface between language and the world. That is what it must do if intentionality – 'aboutness' – is indeed irreducible, as in effect the three ingredients of Putnam's argument combine (if they are sound) to show. To be sure, no aspirant to a purely physicalist version of metaphysical realism could rest content with primitively intentional relations of aboutness. And Putnam may be right to say that "materialism is the only *metaphysical* picture that has contemporary 'clout.' Metaphysics, or the enterprise of describing the 'furniture of the world' … has been rejected by many analytic philosophers … Today, apart from relics, it is only materialists (or 'physicalists,' as they like to call themselves) who continue the traditional enterprise."[35] But it remains to be convincingly explained why "the only sort of metaphysical realism that our time can take seriously" (Putnam, 1989, p. 220) should be a thorough-going physicalism, or why irreducible intentionality should be especially uncomfortable for one of metaphysically realist predilection.

We end with one final reservation about the scope of the argument. If the interpretation offered is sound, then it can engage only a realist who accepts the autonomy of the division of the world into objects and kinds. So far as we can see, it must therefore fail to touch an intermediate, apparently coherent combination of views: the combination which yokes rejection of the idea that there is a 'pre-sliced,' 'ready-made world' – that the world divides into kinds of thing, stuff, and so on quite independently of our efforts to devise a conceptual scheme in terms of which it may be best described and understood – with acceptance of an evidentially unconstrained conception of truth, that is, with realism in the sense Dummett has made familiar (see Chapter 20, REALISM AND ITS OPPOSITIONS). Putnam has sometimes written as if the latter form of realism must fall to his argument. If we are right, that can be so only if a Dumettian, realist conception of truth must in the end consist in the kind of robust correspondence conception which is the essence of metaphysical realism as Putnam conceives it. However, it is one thing to accept that questions about what words refer to make sense, and have determinate answers, only within a conceptual scheme (so that the words cannot be thought of as having reference to an antecedently determinate world of objects and kinds), and another to claim that we cannot combine those words into statements which may, in principle, possess determinate but undetectable truth-values. If the latter *is* a consequence of the former, further argument is needed to show it.[36,37]

Appendix: Permutation Results

In his Appendix to *Reason, Truth and History*, Putnam shows how to prove a relatively strong permutation result to the effect that, given an interpretation I of a (first-order) language L, we can construct another ('unintended') interpretation J which preserves the truth-conditions of all the sentences of L (in his sense, under which sentences have the same truth-condition if they have the same truth-value at all possible worlds), whilst varying the extensions of terms and predicates. Here, we first prove a more basic, weaker result

(to the effect that, given an interpretation of a first-order language, we can always construct an alternative 'unintended' interpretation which coincides with the given interpretation over the truth-values of all the sentences, while varying the extensions of terms and predicates). We then indicate how the method of proof (which differs somewhat from that employed by Putnam) may be extended to obtain, first, a result essentially the same as Putnam's and then some stronger results, for second-order languages and for languages with modal operators.

Weak Permutation

For this, we work with a first-order language L, with logical constants: ¬, ∧, ∃; terms, comprising individual constants a, b, c, … and variables x, y, z, … ; and predicate constants F, G, H, …. The atomic sentences are just the strings $Ft_1, \ldots, t_n$, consisting of an n-place F followed by n occurrences of individual terms. If A,B are sentences, so are ¬A, A ∧ B, and ∃xA(x), where x is any variable and A(x) comes from some sentence A by replacing one or more occurrences of some one individual constant by occurrences of x.

An interpretation I of L consists of a non-empty domain D with assignments of elements of D as denotations of the individual terms and of sets of ordered n-tuples of elements of D, for appropriate choices of n, as extensions of then-place predicates. Thus to each 1-place predicate, I assigns as extension a subset of D – intuitively, the set of elements of D having the property for which, under I, that predicate is taken to stand; to each 2-place predicate, I assigns a set of ordered pairs of elements of D – intuitively, the pairs of elements of D the first of which bears to the second the relation for which, under I, that predicate stands, and so on. 'I(A) = 1' denotes that A is true under I, and is defined as follows:

$I(Ft_1 \ldots t_n) = 1$ iff $\langle I(t_1) \ldots I(t_n) \rangle \in I(F)$

$I(\neg B) = 1$ iff $I(B) \neq 1$

$I(B \wedge C) = 1$ iff $I(B) = I(C) = 1$

$I(\exists xB(x)) = 1$ iff there is an interpretation I° which differs from I at most in its assignment to x, such that $I^{\circ}(B(x)) = 1$

Theorem 1 (weak permutation)

> Let I be any interpretation with domain D, and ϕ be any permutation of D. Let I* be any interpretation with the same domain D such that, for every term t, $I^*(t) = \phi(I(t))$, and for every n-place F, $I^2(F) = \{\langle d_1, \ldots, d_n \rangle \mid \langle (\phi^{-1}(d_1), \ldots, \phi^{-1}(d_n) \rangle \in I(F)\}$. Then for any A, $I(A) = 1 \leftrightarrow 1^*(A) = 1$

Strictly, for the purposes of Putnam-type arguments, we need only establish that for given I, D, and ϕ, there is at least one interpretation meeting the specified conditions on I*, for which the theorem's consequent holds. However, the proof can proceed more smoothly for the theorem as stated. It is obvious that there *are* (non-trivial) interpretations meeting the antecedent conditions.

Proof is by induction on the degree of A, as measured by the number of logical operators occurring in it. So the induction hypothesis (IH) is that the theorem holds for all wffs of

degree < A, and on this hypothesis it is to be proved that the theorem holds for A. More fully stated, IH is:

> If I^1 and I^2 are *any* interpretations with the same domain, such that for each term t, $I^2(t)=\phi(I^1(t))$ and for any n-place F, $I^2(F) = \{\langle d_1, \ldots, d_n \rangle \mid \langle \phi^{-1}(d_1), \ldots, \phi^{-1}(d_n) \rangle \in I(F)\}$, then for any B of degree < A, $I^2(B) = 1 \leftrightarrow I^1(B) = 1$

A is atomic, that is, $Ft_1 \ldots t_n$ for some n.

$$
\begin{aligned}
I^*(Ft_1 \ldots t_n) = 1 \quad \text{iff} \quad & \langle I^*(t_1) \ldots I^*(t_n) \rangle \in I^*(F) \\
\text{iff} \quad & \langle \phi(I(t_1)) \ldots \phi(I(t_n)) \rangle \in I^*(F) \\
\text{iff} \quad & \langle \phi^{-1}(\phi(I(t_1))) \ldots \phi^{-1}(\phi(I(t_n))) \rangle \in I(F) \\
\text{iff} \quad & \langle I(t_1) \ldots I(t_n) \rangle \in I(F) \\
\text{iff} \quad & \langle I(Ft_1 \ldots t_n) \rangle = 1
\end{aligned}
$$

Induction step for ∃

> Suppose $I(\exists xB(x)) = 1$. Then for some I° differing from I in at most its assignment to x, $I°(B(x)) = 1$. Let $I^\#$ be the same as I^* except possibly over its assignment to x, where $I^\#(x) = \phi(I°(x))$. It is easily verified that I° and $I^\#$ meet the conditions on I^1 and I^2 in the induction hypothesis, which then yields that $I^\#(B(x)) = 1$. Hence $I^*(\exists xB(x)) = 1$. The steps are obviously reversible.
>
> Other cases for induction are straightforward. ⊡

Theorem 1 ensures that given any assignment of truth-values to the sentences of L, induced by an interpretation I, there will be a quite different interpretation I^* of L based on a permutation of I's domain, which induces all the same truth-*values* on L's sentences, but makes quite different assignments to the names and predicates of the language.

Strong Permutation

A stronger permutation result will be that given an interpretation I of L, we can get a different interpretation I^* that departs from I over its assignments to names and predicates, whilst giving L's sentences the same truth-*conditions* (in the Putnam sense – the sentences of L coincide in truth-value not just at the actual world, but at every possible world, under the two interpretations). To state and prove this stronger result, we need some preliminary stage-setting:

By a world structure we mean a triple ⟨D,W,σ⟩, where D and W are non-empty sets (intuitively, think of W as the set of all possible worlds, of D as a very inclusive set of objects, containing each object which exists at any of the worlds in W), and σ is a function from W into the non-empty subsets of D (i.e., σ assigns a non-empty subset of objects to each world[38]).

Interpretations I are now assignments as follows: for each i and j, I assigns to the term t_i an element of the domain of w_j as its denotation relative to that world, that is, $I(t_i,w_j) \in \sigma(w_j)$.

And to each n-place F, I assigns, relative to each world w_j, a set of ordered n-tuples from the domain of w_j, that is, $I(F,w_j) \subseteq (\sigma(w_j))^n$.

Truth under I is now of course a relation between sentences of L and worlds, defined thus:

Atoms $I(Ft_1 \ldots t_n, w_j) = 1$ iff $\langle I(t_1,w_j) \ldots I(t_n,w_j) \rangle \in I(F,w_j)$

Molecules $I(\neg B.w_j) = 1$ iff $I(B,w_j) \neq 1$

$I(B \wedge C, w_j) = 1$ iff $I(B,w_j) = 1$ and $I(C,w_j) = 1$

$I(\exists xB(x), w_j) = 1$ iff there is an interpretation I° which differs from I at most in its assignment to x, such that $I^\circ(B(x), w_j) = 1$

Theorem 2 (strong permutation)

Let I be an interpretation of L. Let the ϕ_j be permutations[39] respectively of each of $\sigma(w_j)$ for all the $w_j \in W$. Let I* be any interpretation of L such that for all i and j, $I^*(t_i,w_j) = \phi_j[I(t_i,w_j)]$ and for every n-place F, $I^*(F, w_j) = \{\langle d_1, \ldots, d_n \rangle \mid \langle (\phi_j^{-1}(d_1), \ldots, \phi_j^{-1}(d_n) \rangle \in I(F, w_j)\}$. Then $I^*(A, w_j) = 1 \leftrightarrow I(A, w_j) = 1$

Proof is again by induction on the degree of A – the foregoing proof of *weak permutation* is readily adapted to show what's required for arbitrary w_j, simply by writing in w_j as an extra parameter as appropriate.

A is atomic, that is, $Ft_1 \ldots t_n$ for some n:

$$
\begin{aligned}
I^*(Ft_1,\ldots,t_n,w_j) = 1 \quad \text{iff} \quad & \langle I^*(t_1,w_j),\ldots,I^*(t_n,w_j) \rangle \in I^*(F,w_j) \\
\text{iff} \quad & \langle \phi_j(I(t_1,w_j)),\ldots,\phi_j(I(t_n,w_j)) \rangle \in I^*(F,w_j) \\
\text{iff} \quad & \langle \phi_j^{-1}(\phi_j(I(t_1,w_j))),\ldots,\phi_j^{-1}(\phi_j(I(t_n,w_j))) \rangle \in I(F,w_j) \\
\text{iff} \quad & \langle I(t_1,w_j),\ldots,I(t_n,w_j) \rangle \in I(F,w_j) \\
\text{iff} \quad & I(Ft_1,\ldots,t_n,w_j) = 1
\end{aligned}
$$

As before, the induction step is quite straightforward. Here, for illustration, is the case for $\wedge$:

Suppose $I(B \wedge C, w_j) = 1$. Then $I(B,w_j) = I(C,w_j) = 1$. By IH, $I^*(B,w_j) = I^*(C,w_j) = 1$. Hence $I^*(B \wedge C, w_j) = 1$. Steps obviously reversible.

Strengthening for Second-Order Languages

We extend our first-order language L by permitting binding of (first-level) predicate variables by the second-order existential quantifier ∃f – we use f,g, … as predicate variables. An interpretation of our second-order language L^2 will make assignments to them of entities of the same types as are assigned to predicate constants in the first-order case. 'true under I' is defined as for previous cases, except that we add a clause for the second-order quantifier:

$I(\exists fB(f))=1$ iff there is an interpretation I° which differs from I at most in its assignment to f, such that $I^\circ(B(f)) = 1$

With this addition, we can straightforwardly extend the weak and strong permutation results to the second-order case – all that is needed is an extra case in the induction, dealing with sentences in which the principal operator is second-order ∃. For the second-order extension of Theorem 1, this runs:

Induction step for second-order ∃

Suppose $I(\exists fB(f)) = 1$. Then for some I° differing from I in at most its assignment to f, $I°(B(f)) = 1$. Let $I^{\#}$ be the same as I^{*} except possibly over its assignment to f, where $I^{\#}(f) = \{\langle d_1,\ldots,d_n\rangle \in D^n \mid \langle \phi^{-1}(d_1),\ldots,\phi^{-1}(d_n)\rangle \in I°(f)\}$. Then by the induction hypothesis, $I^{\#}(B(f)) = 1$. Hence $I^{*}(\exists fB(f)) = 1$. The steps are obviously reversible.

Languages with Modal Operators

The addition of a modal operator, say □, to L (or L^2) permits the formation of complex sentences which are not truth-functions of their atomic constituents. That is, we can form sentences B with atomic constituents $A_1 \ldots A_k$ so that B's truth-value at a world w_j is not a function simply of the values of $A_1 \ldots A_k$ at w_j. B's truth-value at w_j is, rather, a function of the values of $A_1 \ldots A_k$ at the other worlds in W.

Does this prevent us from running the permutation argument? Well, it seems that it should *not* do so – just because, while a modal sentence's truth-value at a given world is not a function of the values of its atomic ingredients at *that* world, it is a function of their values at other worlds. But we know from *strong permutation* that we can jiggle the assignments to individual constants and predicates in such a way as to obtain an 'unintended' interpretation which agrees with the original interpretation on the truth-values of all the sentences of L (or L^2) at all possible worlds (so that they have the same truth-conditions, in Putnam's sense). It follows from this that adding □ to L (or L^2), with the usual clause to the effect that $I(\Box B, w_j) = 1$ iff $I(B, w_k) = 1$ for all w_k accessible from w_j, can make no essential difference to the situation. The essential point is this. Given an interpretation I which induces a *pattern* of truth-values on a sentence B across the possible worlds, we can construct a variant interpretation I^{*}, differing from I in its assignments to terms and predicates (and in case of L^2, predicate variables) at those worlds, but agreeing with I on the induced value of B at each world. And that is enough to ensure that I and I^{*} will not diverge over the truth-values of modal functions of B.

Notes

1 "Models and reality," in Putnam (1983, p. 13).

2 Putnam (1981, p. 49): cf. also "A defense of internal realism," in Putnam (1990a, p. 30).

3 The Twin Earth argument was first presented in Putnam's "The meaning of 'meaning,'" see especially Putnam (1975, pp. 223 ff.). An abbreviated statement of it is given in Putnam (1981, pp. 22–29). See also pp. 41–43 for the distinction between pure and impure mental states; and Chapter 8, A GUIDE TO NATURALIZING SEMANTICS.

4 By saying that an assignment of truth-values to sentences meets operational constraints, Putnam means, roughly speaking, that it accords with all the observational data that is available in principle. By theoretical constraints he means whatever further methodological constraints – including pragmatic considerations such as simplicity and economy – guide the optimum choice between theories which meet all operational constraints. Cf. "Models and reality," in Putnam (1983, pp. 3–6).

5 See Putnam's classic characterization of the 'internalist perspective' (1981, pp. 49 ff.).

6 Thus he writes:

> For an internalist like myself, the situation is quite different…. Signs do not intrinsically correspond to objects, independently of how those signs are employed … 'Objects' do not exist independently of conceptual schemes. *We* cut up the world into objects when we introduce one or another scheme of description. Since the objects *and* the signs are alike *internal* to the scheme of description, it is possible to say what matches what…. Indeed, it is trivial to say what any word refers to within the language the word belongs to, by using the word itself. What does 'rabbit' refer to? Why, to rabbits of course. (1981, p. 52)

See also "Models and reality," in Putnam (1983, p. 24).

7 This question has exercised some of Putnam's critics, e.g., Blackburn (1994, p. 27), but needlessly, if we are right.

8 See, e.g., remarks at pp. 25, 27, and 29.

9 Putnam (1981, p. 52). Cf. also:

> if the received view is correct, then we would have an elegant *account of how* intensions and extensions are fixed. [p. 32. our emphasis] One might say that … my 'mental representations' … *refer* to cathood … this may be true, but it just repeats that reference is fixed in one way rather than another. This is what we want to explain and not the explanation sought. [p. 37] To explain reference in terms of (impure) intention would be circular. And the problem of how *pure* mental states of intending, believing, etc., can … constitute reference is just what we have found so puzzling. [p. 43]

10 More about this matter below.

11 Cf. Putnam (1981, p. 34). Putnam's stipulation for (c)-worlds is a little odd – it would have sufficed to have 'cat' refer to cats and 'mat' refer to mats in this case, since all that's required is that 'A cat is on a mat' be false in (c)-worlds.

12 For formal details, see the Appendix.

13 That said, it's worth observing that, even if Putnam's project were to argue for the indeterminacy of reference *tout court*, it's not clear that the permutation argument would be vulnerable to the stated objection. For the proof of the permutability of reference – illustrations apart – is *entirely general*, and following it need involve consideration of no specific suppositions about the reference of particular expressions in the language: suppositions whose status might then be settled by stipulation. Someone – not Putnam – who wanted to harness the permutation argument to a general skepticism about reference *could* quite coherently carry its conclusion forward in the form of the counterfactual: if there were such a thing as determinate reference, it would not be recoverable from the truth-conditions of sentences. And indeed, the overall strategy of arguing for indeterminacy by establishing enough such counterfactuals, with a sufficient variety of consequents ('…, it would not be recoverable from speakers' intentions,' '…, it would not be recoverable from facts about causality,' etc.), is a perfectly coherent one. By the same token, though, the concern – for a supporter of Putnam – that the model-theoretic argument may fail stably to focus against metaphysical realism, dissolving instead into 'Putnam's Paradox,' is not so easily set aside.

14 See, for instance, Ian Hacking (1983, p. 105); though this may not be quite fair to Hacking, who in the relevant passage is mainly raising a doubt about the first-order formalizability, e.g., of physical theory, and is not really emphasizing the failure of the Lowenheim–Skolem Theorem at second order. Cf. Putnam's remarks in n. 11 of Putnam (1989, p. 230).

15 Permutation results for second-order languages and languages with the usual modal operators are outlined in the Appendix.

16 There are, however, reasons to qualify the force of this reservation, whose significance will emerge only when more has been done to explain how Putnam's argument can bear selectively on metaphysical realism. See §VII, and especially n. 31 below.

17 Hartry Field (1972). Field's view is discussed by Putnam (1981, pp. 45–46; 1978, pp. 14–17, 30–32, and 57–58).

18 A reason for thinking the tension merely apparent will anyway emerge in §VI below – see also n. 25.

19 Besides directing it at Field's in the passage quoted, Putnam (1983, p. 18) makes essentially the same move against Evans's (1973) version of the causal theory, and (1989, pp. 219–220) against Devitt's appeal to a causal theory (q.v. Devitt, 1983).

20 This claim appears to run counter to Putnam's own view, as expressed in "Model theory and the 'factuality' of semantics" (1989). He stresses there that his model-theoretic argument is directed against a limited target – *physicalistic* metaphysical realism. Certainly some of the argumentation rehearsed in that paper relies upon the assumption that the metaphysical realist aspires to a physicalist account of reality – including the circularity argument discussed above. Our point is that the 'just more theory' move is *not* subject to this limitation. A similar point is made by David Lewis (1984, pp. 232–233).

21 Something like this may well be the intended thrust of Putnam's complaint (1983, p. xi) that the causal realist "ignores his own epistemological position."

22 Putnam advances two quite distinct objections against the proposal. This is his first objection; we shall discuss the second in due course. Meanwhile, note that the first objection is to any identification of reference with a physicalistic relation, regardless of whether it is made in the interests of defending metaphysical realism.

23 Putnam (1989, p. 217). Putnam has 'requirement' where we have 'proposal.' The requirement to which he refers is presumably that if a relation R is to be the 'intended' reference relation, the supposition that R is the reference relation should yield an explanation of facts about our use of words.

24 A more detailed formulation of this argument is given in "Beyond historicism" (1983, pp. 290–298). We lack space to discuss it here, but it seems to us that it is vitiated by the same gratuitous assumption that anyone who proposes a 'theoretical identification' of an intentional notion – such as *explanation* or *reference* – is thereby debarred from using the notion in question in arguing for the identification. That this might be a reasonable restriction to impose on attempts at *analytic* reductions of intentional notions seems quite irrelevant.

25 Cf. "Models and reality" in Putnam (1983, p. 17). This bears on the interpretative question left dangling in §IV – it is clearly quite consistent with holding that a causal constraint needs to be met in many (or even all) cases of genuine reference to deny that a full constitutive account of reference may be given in purely causal terms.

26 Here is a relevant passage from "Model theory and the 'factuality' of semantics" at p. 219:

> if *E(T)* is the event of someone's using a token of a term *T*, then there is a good sense of 'causal connection' in which *every* event in the backward light-cone of *E(T)* is 'causally connected' to the event *E(T)*; but it will almost never be the case that the term *T*… *refers* to every event in the backward light-cone of *E(T)* (and it will typically be the case that the term does refer to things with which the token is *not* causally connected, e.g., future things).

27 "Models and reality" in Putnam (1983, p. 18). See Evans (1973, pp. 187–208) for his version of the theory. For some further discussion of this approach, see Chapter 35, REFERENCE AND NECESSITY, esp. §4.

28 Putnam would probably concede that the first difficulty may not be insuperable – cf. his acknowledgement that Evans has a proposal, to which he offers no objection – to deal with this problem: see the footnote on p. 18 of "Models and reality," in Putnam (1983). He does press the difficulty a little further, claiming that the distinction between causes and background conditions is inescapably interest-relative; but this shows at best that the relevant causal relations can't be singled out by appeal to that distinction, not that they can't be singled out at all.

29 Further objections, complementing those brought by Putnam (and those discussed in Chapter 8, A GUIDE TO NATURALIZING SEMANTICS), and including a forceful play with the holism of the mental, are developed by Paul Boghossian (1991).

30 Putnam, of course, is well aware of the possibility of this response to his argument, envisaging it explicitly (1989, p. 220); however, he does not regard acknowledging the primitiveness of reference

as a commitment to regarding it as 'simple and irreducible.' *Representation and Reality* is, in effect, an extended argument to the contrary.

31 The astute reader will note that if these considerations are indeed the key to the question of how Putnam's argument can tell selectively against metaphysical realism, then there actually is little force – in the resultant dialectical setting – in the reservation with which our discussion of the significance of the permutation argument concluded in §III. That reservation was, in effect, that while permutation-based reinterpretations of a language might be *consistent* with all data concerning the use of its sentences, they would be likely to be dominated by the preferred interpretation once appropriate constraints on the construction of interpretational theories, beyond adequacy to the linguistic data, are allowed their proper influence. But that, if correct, is a point about the methodology of interpretation – something which can be freely acknowledged from the internal realist point of view as conditioning the concept of truth that applies to ascriptions of reference and other semantic claims, but which takes us no closer to the *constitutive* account which the metaphysical realist needs of the nature of reference, conceived as a network of external relations of which methodologically superior interpretation is, at best, a means of discovery.

32 Recall the complaint which Putnam airs against Lewis's positive view, that it amounts to "saying that we-know-not-what fixes the reference relation we-know-not-how" (1989, p. 220).

33 So far as we are aware. Putnam does not himself explicitly employ the metaphor of 'slicing.' But it is common in discussion of his ideas and implicit in several of his own characterizations of internal realism. For example, "'Objects' do not exist independently of conceptual schemes. *We* cut up the world into objects when we introduce one or another scheme of description" (Putnam, 1981, p. 52). Of a piece with this are his frequent characterizations of external or metaphysical realism as involving – via its commitment to the idea that there is, whether we can discover it or not, just one true theory of the world – a belief that there is a 'ready-made world,' having an intrinsic or 'built-in' structure, comprising a 'fixed totality of mind-independent objects.' Cf., for example, Putnam (1983, p. 211; 1981, p. 49).

34 But an outright repudiation of the idea of sortal predetermination, even if not accompanied by a lurch into constructivist metaphor, would be in at least *prima facie* conflict with retention of the idea, of which Putnam himself has been a principal advocate, that the world encompasses various *natural kinds*. The apparent tension here runs parallel to that noted earlier, between Putnam's advocacy of an externalist account, in broadly causal terms, of how reference is 'fixed,' on the one hand, and on the other, his insistence that no progress can be made on the problem of explaining how reference can be determinate by appeal to causal relations between our words and appropriate bits of the world. So, unless the tension can be argued to be merely apparent, some qualification is needed. We cannot pursue this somewhat delicate issue here, and must content ourselves with one brief cautionary remark. Even supposing that the repudiation of sortal predetermination needs qualification to make space for belief in natural kinds, it would be a mistake, for at least two reasons, to think that this could be exploited to recover a metaphysically realist conception of determinate reference. First, the hypothesis that certain things instantiate a natural kind would, at best, serve to explain the *unity* of a class of things forming the reference or extension of a predicate as distinct from explaining what *constitutes* reference to that class. (This point may, we suspect, contain the germ of a resolution of the apparent tension – but that is a further issue.) Second, however precisely the envisaged qualification might run, it would be restricted in scope in a way which would, even prescinding from the previous point, preclude its yielding a fully general solution to the problem with which Putnam confronts the metaphysical realist. Crucially, we could expect no help with explaining how non-natural-kind terms can enjoy determinate reference. Essentially the same limitation vitiates David Lewis's proposal (1984, pp. 226–229) that some things, such as rabbits, are more eligible to be the referents of our words than others, such as undetached rabbit parts, and that rival schemes of reference may be ranked as better or worse to the extent that their assignments of referents respect 'nature's joints.' Indeed, the difficulty is not just that appeal to natural divisions could afford an at best partial solution to the general problem; it can be seen, on reflection, that it fails to accomplish even that much – the permutation

argument can just as well work to deliver perverse jigglings of perfectly *eligible* referents, and has no need for play with unnatural divisions at all.

35 From "Why there isn't a ready-made world" in Putnam (1983, p. 208).

36 Indeed, Putnam himself has recently shown signs of a cooling in his opposition to realism as Dummett conceives it (e.g., 1994, pp. 503, 510–511).

37 We are indebted to Philip Percival, and to colleagues who attended the Putnam conference in Utrecht in September 1994, especially Putnam himself.

38 The point of this complication is simply to avoid making the needlessly restrictive – and unrealistic – assumption that possible worlds do not differ in point of which objects they contain. In the special case where that assumption holds, we could dispense with the function σ, and need only consider a single permutation ϕ of the domain common to all possible worlds. This special case is, of course, covered by Theorem 2 as stated.

39 Each of the permutations ϕ_j could, of course, be defined to be the restriction to $\sigma(w_j)$ of a single permutation ϕ of the inclusive set D.

References

Blackburn, S. 1994. "Enchanting views." In *Reading Putnam*, edited by P. Clark and B. Hale, pp. 12–30. Oxford: Blackwell.

Boghossian, P. 1991. "Naturalizing content." In *Meaning in Mind: Fodor and his Critics*, edited by B. Loewer and G. Rey, pp. 65–86. Oxford: Blackwell.

Devitt, M. 1983. "Realism and the renegade Putnam." *Nous*, 17(2): 291–301.

Evans, G. 1973. "The causal theory of names." *Aristotelian Society*, suppl. vol. 47: 187–208.

Field, H. 1972. "Tarski's theory of truth." *Journal of Philosophy*, 69(13): 347–375.

Fodor, J. 1990. *A Theory of Content*. Cambridge, MA: MIT Press.

Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Lewis, D. 1984. "Putnam's paradox." *Australasian Journal of Philosophy*, 62(3): 221–236.

Putnam, H. 1975. "The meaning of 'meaning.'" In *Mind, Language and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.

Putnam, H. 1977. "Models and reality." In Putnam, 1983.

Putnam, H. 1978. *Meaning and the Moral Sciences*. Boston and London: Routledge and Kegan Paul.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Putnam, H. 1983. *Realism and Reason: Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press.

Putnam, H. 1989. "Model theory and the 'factuality' of semantics." In *Reflections on Chomsky*, edited by A. George, pp. 213–232. Oxford: Blackwell.

Putnam, H. 1990a. "A defense of internal realism." In Putnam, 1990b.

Putnam, H. 1990b. *Realism with a Human Face*. Cambridge, MA: Harvard University Press.

Putnam, H. 1992. *Renewing Philosophy*. Cambridge, MA: Harvard University Press (based on the Gifford Lectures given at St Andrews in 1990).

Putnam, H. 1994. "Sense, nonsense and the senses: an enquiry into the powers of the human mind." *Journal of Philosophy*, 91(9): 445–517 (based on the John Dewey Lectures given at Columbia University in 1994).

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Further Reading

Brueckner, A. 1984. "Putnam's model-theoretic argument against metaphysical realism." *Analysis*, 44(3): 134–140.

Putnam, H. 1981. "Beyond historicism." In Putnam, 1983.
Putnam, H. 1981. "Why there isn't a ready-made world." In Putnam, 1983.
Putnam, H. 1988. *Representation and Reality*. Cambridge, MA: Bradford/MIT Press.
Putnam, H. 1994. "Simon Blackburn on internal realism." In *Reading Putnam*, edited by P. Clark and B. Hale, pp. 242–254. Oxford: Blackwell.

Postscript: Recent Work on Putnam's Model-Theoretic Argument

JUSSI HAUKIOJA

During the past 15–20 years, Putnam's model-theoretic argument has no longer been a hot and fashionable topic in philosophy. It has, however, secured its place as one of three central arguments in analytic philosophy where *referential indeterminacy* figures in a central role, the other two being Quine's argument for referential inscrutability (see Chapter 26, INDETERMINACY OF TRANSLATION), and Kripke's Wittgenstein's skeptical paradox about rules (see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING). Nonetheless, interest in features particular to the model-theoretic argument, its proper interpretation, and its role in Putnam's argumentation as well as in anti-realist arguments in general, has by no means disappeared. A number of articles putting forward interesting new suggestions in these regards have come out, as well as two books (Taylor, 2006; Button, 2013) that build centrally on Putnam's argument, or argumentation directly inspired by it.

In this postscript, I want to highlight two strands in the recent discussions.[1] Both of these can be seen as attempts to bring new life to two ideas that had come under heavy criticism in the earlier debates: the 'just more theory' maneuver that Putnam employs as a part of the model-theoretic argument, and referential magnetism as a way of *responding* to Putnam's arguments.

1 Reconsidering the 'Just More Theory' Maneuver

In discussions of Putnam's model-theoretic argument, a near consensus seemed to have been reached in the 1990s that the 'just more theory' maneuver (henceforth the 'JMT maneuver') is question begging (cf. Hale and Wright, §V). It is one thing to say that, for example, causation singles out the intended interpretation for our theories, and another to say that *causation-talk* does that. The metaphysical realist who appeals to causation intends, of course, the former but not the latter. Therefore, to deny the realist the possibility to appeal to causal constraints in explaining word-world relations – including the word-world relations for terms such as 'causation' – is to beg the question. The JMT maneuver, according to this response, is based on a fairly elementary use-mention confusion.

Some commentators have recently expressed doubts about this consensus, and gone on to present reconstructions of the model-theoretic argument in an attempt to show that there is more to Putnam's maneuver than first meets the eye. A helpful starting point to this is in fact found already in some of Lewis's comments on Putnam (Lewis, 1984, p. 233). The JMT maneuver was clearly intended to work not just against causal constraints, but against *any* constraints that the metaphysical realist could come up with. But if the JMT maneuver were intended as a *completely* general one, we should expect it to work just as well against

"magical," non-naturalistic constraints as against naturalistic ones. But, since Putnam (mockingly) introduces a "magical theory of reference" as the metaphysical realist's only recourse, clearly he thought such a view would be immune to his argument. To be charitable to Putnam, we should at least try to see whether there is something specific to the metaphysical realist's position, and the constraints available to it, to make it vulnerable to the JMT maneuver.

Igor Douven (1999) and Tim Button (2013) have sought to reconstruct the model-theoretic argument in ways that define the intended target[2] of the argument more clearly, as a certain sort of meta-semantic naturalism.[3] According to meta-semantic naturalism, our theory of what singles out an intended interpretation for our words is an empirical theory, completely on a par with other scientific theories. Putnam is assuming that any metaphysical realist will be committed to meta-semantic naturalism, and it is *this* assumption that makes metaphysical realists vulnerable to the JMT maneuver. According to metaphysical realism, as defined by Putnam, an epistemically ideal theory may be false. Given meta-semantic naturalism, an epistemically ideal *meta-semantic* theory may be false: consequently the meta-semantic constraints are 'just more theory' *by the realist's own lights* (cf. Douven, 1999, pp. 488–490).[4]

Button (2013) presents the JMT maneuver as a dilemma posed to the metaphysical realist. Either the realist's proposed attempts to constrain reference have empirical content, or they do not. If they do, they are 'just more theory.' If they do not, then the realist is really invoking something like magic, and not living up to his or her own meta-philosophical ideals. Button goes on to argue, in considerable detail, that although the primary target of Putnam's model-theoretic arguments is a "bracketed empiricist" view – a version of logical positivism, or a direct descendant – the arguments can in fact be used against a variety of present-day varieties of metaphysical realism. The result, according to Button, is that metaphysical realists (or "external" realists, in Button's terminology) who accept that skeptical doubts about the external world make sense, cannot stop a slide from Cartesian skepticism to *Kantian* skepticism, questioning our very ability to represent an external world. However, Kantian skepticism is self-refuting, and we should thus reject the external realism that led to it.

2 Reference Magnetism

Another recent development in discussions of Putnam's model-theoretic argument – as well as of other arguments for meaning skepticism or referential indeterminacy – is the rise of *reference magnetism* as a respectable constraint on interpretation. The central idea, suggested by David Lewis[5] (1984), is that some properties are *reference magnets*, that is, that they are by their very nature more eligible to be the referents of our terms because they are *natural* properties, carving nature at its joints. Lewis's suggestion was long taken to be an *ad hoc* move (or even a piece of "occult metaphysics"; cf. Sider, 2011, p. 27), not to be taken seriously. In recent years, however, many theorists have sought to make the proposal seem more respectable.

The revival started (at least in print) with Weatherson (2003), where it was suggested that the classical analysis of knowledge could perhaps be defended against Gettier-style counter-examples by appealing to the *naturalness* of the category of justified true belief (or its superior naturalness relative to proposals that better respect our ordinary classifications). Williams (2007) and Sider (2011) present a defense of reference magnetism as a

general principle in meta-semantics, attempting to deflect charges of *ad hoc*-ness by deriving it from general doctrines about theoretical virtue. Williams and Sider argue that reference is an explanatory notion, and explanatory properties and relations should, in general, be cast in joint-carving terms: therefore, we have good reason to expect referential relations to be natural (or at least reasonably so).

There are two potential problems with this suggestion. First, it is one thing to say that the *reference relation* should carve nature at its joints, and quite another to say that the properties *connected to our terms* via the reference relation do so. The former may follow from general principles about explanatory virtue, but the reference magnetist is not putting forward the former claim, but rather the latter. And it is not obvious at all why the former should entail, or even give justification, for the latter: why should a joint-carving reference relation relate our terms to natural, rather than gerrymandered properties?[6]

Second, even if we agree that reference is an explanatory notion in semantics, a lot will turn on exactly *which explanatory projects* we expect referential relations to play a central role in. If, as seems plausible, referential relations are centrally involved in the explanation of successful linguistic communication, then it is by no means obvious that reference magnetism adds to the explanatory power of our theories. An example by Schwartz nicely illustrates the second problem: "Suppose a sentence S is systematically used to communicate that a certain state of affairs p obtains: speakers try to utter S if and only if p obtains, hearers take utterances of S as evidence for p, and so on. If magnetism is true, then it is a live possibility that, unbeknownst to everyone, S actually means not p but q, because q concerns objectively more natural properties" (Schwartz, 2014, p. 31).[7]

* * *

It remains to be seen how damaging the new reinterpretations of the model-theoretic argument, presented by Douven and Button, are against present-day versions of metaphysical realism. If they *do* work, then realists will have to renounce their meta-semantic naturalism and admit that meta-semantics is *not* an ordinary empirical enterprise, and try to make this concession somehow consonant with their overall metaphysical realist view.[8] The recourse to reference magnetism can be seen as one way of trying to accomplish this: to say that the choice between empirically equivalent interpretations of the total theory is to be decided by comparisons of relative naturalness is already to admit that the meta-semantics is not determined by empirically available evidence. But there is no reason to think that reference magnetism is the only way of doing this. At the very least, a robustly realist overall position in metaphysics does not seem to depend on a naturalistic conception of meta-semantics – nor is it immediately obvious that a metaphysical realist view as characterized by Putnam, insisting that even empirically ideal theories may be false, could not be combined with a non-naturalist conception of meta-semantics.

Notes

1 Two other strands, which will not be presented in detail here, are Bays's (2001) criticism of Putnam's use of model theory, and Chambers's (2000) claim that Putnam's argument is simply unsound. For criticism and discussion of Bays see Bellotti (2005), Bays (2007), Button (2011); for criticism and discussion of Chambers see Haukioja (2001), Kroon (2001), Chambers (2001).

2 Or rather, especially in Button's case, the proper target of a Putnam-style model-theoretic argument.

3 An earlier (re)interpretation of the model-theoretic argument that is in many ways similar was put forward by Zalabardo (1998): here, the aim of the argument is to show that if the metaphysical realist's conception of how referential facts are determined is right, we would not be able to grasp those very facts.

4 If this is the right way of construing the target of the argument, we can also see the relevant difference between Putnam's argument and Bays's (2008) "astronomical" argument (which Bays claims to be structurally similar to Putnam's): the crucial premise in Putnam's argument, stating that nothing but theoretical or operational constraints can single out the intended interpretation, is a part of meta-semantic naturalism, which Putnam uses as a premise in his *reductio*. The corresponding premise in Bays's "astronomical" argument is not a part of any (sensible) philosophical world view.

5 Schwartz (2014) argues, persuasively, that in fact reference magnetism was not a part of Lewis's account of language (nor was the global descriptivism that he, in responding to Putnam's model-theoretic argument, suggested should be enhanced by reference magnetism), although the naturalness of properties did play a role in blocking unintended interpretations in his theory of mental content.

6 Sider (2011, pp. 28–29) discusses the problem briefly, but concludes, for reasons that remain somewhat unclear, that the problem will not be a serious one if the reference relation is merely "reasonably" rather than "perfectly" joint-carving.

7 For further argument against reference magnetism, along similar lines, see David Chalmers's "Twentieth excursus: reference magnets and the grounds of Intentionality." http://consc.net/books/ctw/excursus20.pdf (accessed October 6, 2016).

8 Another option will, of course, be to reject metaphysical realism. But our problems do not simply disappear if we become anti-realists. Button, for example, finds anti-realism (as well as Putnam's "internal" realism) just as unacceptable as "external" realism, recommending that we, in the end, should steer a middle course between "external" and "internal" realism.

References

Bays, T. 2001. "On Putnam and his models." *Journal of Philosophy*, 98(7): 331–350.

Bays, T. 2007. "More on Putnam's models: a reply to Bellotti." *Erkenntnis*, 67(1): 119–135.

Bays, T. 2008. "Two arguments against realism." *The Philosophical Quarterly*, 58(231): 193–213.

Bellotti, L. 2005. "Putnam and constructibility." *Erkenntnis*, 62(3): 395–409.

Button, T. 2011. "The metamathematics of Putnam's model-theoretic arguments." *Erkenntnis*, 74(3): 321–349.

Button, T. 2013. *The Limits of Realism*. Oxford: Oxford University Press.

Chalmers, D. 2012. *Constructing the World*. Oxford: Oxford University Press.

Chambers, T. 2000. "A quick reply to Putnam's paradox." *Mind*, 109(434): 195–197.

Chambers, T. 2001. "Putnam's paradox: a less quick reply to Haukioja and Kroon." *Mind*, 110(439): 709–714.

Douven, I. 1999. "Putnam's model-theoretic argument reconstructed." *Journal of Philosophy*, 96(9): 479–490.

Haukioja, J. 2001. "Not so quick: a reply to Chambers." *Mind*, 110(439): 699–702.

Kroon, F. 2001. "Chambers on Putnam's paradox." *Mind*, 110(439): 703–708.

Lewis, D. 1984. "Putnam's paradox." *Australasian Journal of Philosophy*, 62(3): 221–236.

Schwartz, W. 2014. "Against magnetism." *Australasian Journal of Philosophy*, 92(1): 17–36.

Sider, T. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Taylor, B. 2006. *Models, Truth, and Realism*. Oxford: Oxford University Press.

Weatherson, B. 2003. "What good are counterexamples?" *Philosophical Studies*, 115(1): 1–31.

Williams, J. R. G. 2007. "Eligibility and inscrutability." *The Philosophical Review*, 116(3): 361–399.

Zalabardo, J. L. 1998. "Putting reference beyond belief." *Philosophical Studies*, 91(3): 221–257.

28

Sorites

MARK SAINSBURY AND TIMOTHY WILLIAMSON

1 The Early History

The logician Eubulides of Miletus, a contemporary of Aristotle, was famous for seven puzzles. One was the Liar: If a man says that he is lying, is he telling the truth? Another was the Hooded Man: How can you know your brother when you do not know that hooded man, who is in fact your brother? There were also the Bald Man and the Heap. In antiquity they were usually formulated as series of questions. Does one grain of wheat make a heap? Do two grains of wheat make a heap? Do 10,000 grains of wheat make a heap? If you admit that one grain does not make a heap, and are unwilling to make a fuss about the addition of any single grain, you are eventually forced to admit that 10,000 grains do not make a heap. Is a man with one hair on his head bald? Is a man with two hairs on his head bald? Is a man with 10,000 hairs on his head bald? If you admit that a man with one hair is bald, and are unwilling to make a fuss about the addition of any single hair, you are eventually forced to admit that a man with 10,000 hairs is bald. The standard ancient terms for the Heap and the Bald Man were "sorites" (from "soros," a heap) and "phalakros," respectively. Later, "sorites" was used for all such puzzles. They were also known as little-by-little arguments.[1]

Many philosophical doctrines have been suggested as the target Eubulides intended his Heap and Bald Man to destroy: the coherence of empirical concepts, the law of non-contradiction, the Law of Excluded Middle, pluralism, Aristotle's theory of infinity or of the mean. The evidence gives little support to any of these suggestions. Eubulides is indeed said to have attacked Aristotle, but in slanderous terms; the sources do not connect the dispute with any of the puzzles. Aristotle betrays no clear awareness of sorites reasoning in any of his extant works. Some later commentators did consider its use against Aristotle's theory of the mean, but without suggesting that either Eubulides or Aristotle had done so. Eubulides's interests were described as purely logical; if he had a specific target in mind, it is likely to have been a logical one.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Sorites puzzles became a standard weapon of Skeptics in their attacks on Stoic philosophy. A Skeptic does not feel obliged to answer any of the sorites questions; he can simply plead ignorance. If a Stoic is obliged to answer each question "Yes" or "No," he will find himself in an embarrassing position. An obvious focus for Skeptical attack was the Stoic theory of knowledge. It was based on cognitive impressions, which represent real objects with complete accuracy and reliability (compare Descartes's clear and distinct ideas). The Skeptics constructed sorites series from cognitive to non-cognitive impressions, replacing each impression by a virtually indistinguishable one, and took themselves to have undermined Stoic claims to knowledge. Stoic defenses against these attacks were mustered by Chrysippus (*c.*280–*c.*207 BCE), the man with the best claim to have initiated propositional logic.

The Stoics firmly accepted the principle of bivalence: every proposition is either true or false. Chrysippus "strained every nerve" to persuade people of it. For any proposition *P* there is one right answer to the question "*P*?" "Yes" or "No"; for any sequence of propositions P_1, …, P_n there is one sequence of right answers to the questions "P_1?" …, "P_n?" The Stoics used "Are *i* few?" as the schematic form of the *i*th sorites question; thus the right answers to the first and last questions are "Yes" and "No" respectively, and there is a last question, "Are *i* few?" rightly answerable "Yes," immediately followed by a first question, "Are *i* + 1 few?" rightly answerable "No": *i* are few and *i* + 1 are not few; *i* is a cut-off point for fewness.

The Stoics distinguished between sentences and the propositions they are used to assert. The argument from bivalence to the existence of a cut-off point assumes that the sentences "One is few," …, "10,000 are few" express propositions. However, someone who utters "*i* are few" with the sense "A man with *i* hairs on his head is bald" does assert something, which on the Stoic view requires the sentence to express a proposition. The assumption gave no escape from the argument for a cut-off point. Indeed, there is independent evidence that the Stoics accepted the conclusion of the argument. In other cases which look susceptible to sorites reasoning they insisted on cut-off points; for example, they denied that there are degrees of virtue, holding that one is either vicious or perfectly virtuous. An analogy was drawn with a drowning man as he rises to the surface; he is coming closer to not drowning, but he is not drowning to a lesser degree until he breaks the surface, when he is suddenly not drowning at all. Moreover, in rebutting the sorites argument against cognitive impressions, Chrysippus dealt explicitly with the case "when the last cognitive impression lies next to the first non-cognitive one": cognitiveness has a cut-off point. The Stoics were prepared to apply bivalence to sorites reasoning and swallow the consequences: any difficulty in answering the sorites questions must come from our ignorance of the right answers, not their non-existence. The puzzle is an epistemological one.

One might answer the questions "Is one few?" … "Are *i* few?" "Yes," and the questions "Are *i* + 1 few?" … "Are 10,000 few?" "No": but one would be guessing. No one has such knowledge of cut-off points; no one knows both that *i* are few and that *i* + 1 are not few. Such a pattern of answers is forbidden by the principle that one should give an answer only if one knows it to be correct. The wise man, the Stoic ideal, conformed to that principle. Since he was infallible rather than omniscient, he would sometimes suspend judgment. The Stoics did not claim to be wise men, still less to be omniscient; they readily admitted that they did not know whether the number of stars was odd or even and that they could not distinguish between very similar hairs or grains (the examples are ancient). Nevertheless, the aim was to avoid error by withholding assent from what one did not know.[2] The Stoic who did not know enough to be wise should suspend judgment more often than the wise

man. That fits Chrysippus's recommended response to the sorites. At some point in the interrogation one should fall silent.

If sorites questions puzzled the Stoic simply because he did not always know whether to answer "Yes" or "No," like the question about the number of stars, he could confidently face the interrogation armed only with the three possible answers "Yes," "No," and "I don't know." If he knew *i* to be few he would answer "Yes"; if he knew *i* not to be few he would answer "No"; in every other case he would answer "I don't know." Why should an honest admission of ignorance not completely dissolve the puzzle?

However, Chrysippus did not say that one should admit ignorance; he said that one should fall silent. Under interrogation, saying "I don't know" is quite a different policy from saying nothing. The former but not the latter denies knowledge. This undermines the argument that the Stoic could answer each question "Yes," "No," or "I don't know."

The Stoic is supposed to say only what he knows to be correct. "I don't know" in answer to "Are *i* few?" is tantamount to the assertion "I don't know that *i* are few and I don't know that *i* are not few," just as "Yes" is tantamount to "*i* are few" and "No" to "*i* are not few." Thus the Stoic is supposed to answer "I don't know" only if he *knows* that he doesn't know whether *i* are few. The "Yes"/"No"/"Don't know" strategy requires the Stoic to answer "I don't know" whenever he doesn't know. It is therefore available, on Stoic terms, only if whenever one doesn't know whether *i* are few, one knows that one doesn't know whether *i* are few. For simplicity, one may be assumed to know a proposition just in case it is clear, where the logical consequences of what is clear are themselves clear. The prerequisite for the "Yes"/"No"/"Don't know" strategy is then that if *i* are neither clearly few nor clearly not few, *i* are clearly neither clearly few nor clearly not few. This is equivalent on Stoic terms to a pair of simpler principles:

(1a) If *i* are not clearly few, *i* are clearly not clearly few.
(1b) If *i* are not clearly not few, *i* are clearly not clearly not few.

Principles (1a) and (1b) are simply the relevant instances of the "S5" principle for clarity: if something is not clearly so, it is clearly not clearly so. Thus the "Yes"/"No"/"Don't know" strategy is available only if the S5 principle applies. However, the S5 principle is incorrect for clarity in such cases, for clear fewness is as sorites-susceptible as fewness. One is clearly few; 10,000 are not clearly few. By Stoic logic, there is a cut-off point for clear fewness: for some i, $i - 1$ are clearly few and i are not clearly few. Where that point comes is no clearer for clear fewness than for fewness. It is very slightly clearer that $i - 1$ are few than that i are few. But i are too close to being clearly few to be *clearly* not clearly few. One cannot reliably judge whether i are clearly few. In particular, one cannot answer the question "Are *i* clearly few?" just by following the policy: if you hesitate to say "Yes," say "No." If that policy worked, one would unhesitatingly judge that *i* were clearly few if and only if *i* were clearly few; whatever one thought was right would be right. But unless one is reasonably cautious, one will sometimes unhesitatingly judge that *i* are clearly few when they are not in fact clearly few (or even few). Most of us say silly things from time to time. On the other hand, if one is reasonably cautious, one will sometimes hesitate over what turns out to be genuinely clear, for fear of a hidden catch. Principle (1a) is false for some *i*. Principle (1b) fails similarly. Silence is the best policy.

When should the Stoic fall silent? Chrysippus seems to have advised that one should fall silent *before* the end of the clear cases. Stoic constraints may imply that one should

sometimes give no answer to the sorites question "Are *i* few?" even though *i* are clearly few. If one answers "Yes" whenever that answer is clearly correct, on Stoic assumptions one stops answering "Yes" either when it ceases to be clearly correct or later. In the former case one has located the cut-off point for clarity with perfect accuracy; in the latter one has violated the constraint that all one's answers should be clearly correct. Given the failure of the S5 principle for clarity, one cannot reliably locate the cut-off point for clarity with perfect accuracy; thus one will reliably satisfy the constraint that one's answers should be clearly correct only if one stops answering "Yes" before it has ceased to be the clearly correct answer. One must undershoot in order to avoid the risk of overshooting.

The point generalizes. One would like to satisfy two conditions:

(2a) If "Yes" is a good answer, say "Yes."
(2b) If "Yes" isn't a good answer, don't say "Yes."

The goodness of an answer is some truth-related property of it, and does not simply consist in its being given. There is play between the antecedents and consequents of (2a) and (2b); in an imperfect world they will sometimes come apart. In such a case, one either fails to say "Yes" when "Yes" is a good answer, violating (2a), or says "Yes" when "Yes" is not a good answer, violating (2b). If one regards violations of (2a) and (2b) as equally serious, one may simply aim to say "Yes" when and only when it is a good answer. Other things being equal, one's misses are as likely to fall on one side of the target as on the other, and no matter. But one might regard a violation of (2b) as worse than a violation of (2a); one would rather commit an error of omission by not saying "Yes" when it is a good answer than one of commission by saying "Yes" when it is not a good answer. For example, one might prefer failing to make true or warranted statements to making false or unwarranted ones, and follow a policy of saying nothing when in doubt. One decreases the risk of more serious violations by increasing the risk of less serious ones. At the limit, the price of never violating (2b) is sometimes violating (2a). That is the choice the Stoic makes in falling silent before the end of the clear cases, where clarity is goodness. It was worse to say "Yes" in an unclear case than not to say it in a clear one. Those who take the opposite view should fall silent after the end of the clear cases. The Chrysippan strategy can be seen as resulting from two levels of precaution. At the first level, goodness in (2a) and (2b) is simply truth. The Stoics were not alone in holding it to be worse to give a false answer than to fail to give a true one. For truth, (2a) rather than (2b) is to be violated. This preference motivates the constraint that one should give an answer only if it is clear. But then clarity takes on a life of its own as a cognitive end, and again the Stoic takes the cautious option. Condition (2a) rather than (2b) is to be violated for clarity too.

The Skeptics were not satisfied with Chrysippus's silence; it was most notably attacked half a century after his death by Carneades. "For all I care you can snore, not just become quiescent. But what's the point? In time there'll be someone to wake you up and question you in the same fashion." Chrysippus was dialectically no better off than he would have been had he fallen asleep, and Carneades's attitude was that of a chess player with what he takes to be a winning strategy, whose opponent simply refuses to make a move (in a game without time limits).

Suspension of judgment was the Skeptical attitude, and Carneades fastened on the extent to which Chrysippus's strategy allowed it to spread. If Chrysippus suspended judgment in

clear cases, on what basis did he object to the Skeptic's suspension of judgment? The question does not reduce the strategy to immediate incoherence, for some sort of reply is open to Chrysippus: do not suspend judgment when the case is clearly clear. Nevertheless, the Stoics were in a very delicate position. Their epistemological caution enlarged the concessions to Skepticism that their bivalent semantics forced them to make under sorites questioning. The concessions did not amount to surrender, for cases remained in which they could still claim knowledge; but these cases were marked off by a disputed no-man's-land rather than a compelling principle.

So far, the Heap and the Bald Man have been presented, as they usually were in antiquity, as series of questions, not as arguments with premises and conclusions. Yet one speaks of them as paradoxes, and a paradox may be defined as an apparently valid argument with apparently true premises and an apparently false conclusion. In argument form, the sorites goes like this:

Premise 1	1 is few
Premise 2	If 1 is few then 2 are few
Premise 3	If 2 are few then 3 are few
…	
Premise 10, 000	If 9,999 are few then 10,000 are few
Conclusion	10,000 are few

The argument appears to be valid; if its premises are true, its conclusion will be true too. The relevant rule of inference is *modus ponens* (MP), which allows one to infer *Q* from *P* and "If *P* then *Q*"; its validity is hard to challenge. By MP, "1 is few" and "If 1 is few then 2 are few" entail "2 are few." In the same way, "2 are few" and "If 2 are few then 3 are few" entail "3 are few." After 9,999 applications of MP, one reaches the conclusion "10,000 are few." The premise "1 is few" is apparently true and the conclusion "10,000 are few" apparently false. The gradualness of the sorites series makes each of the conditional premises appear true. Thus the apparently valid argument has apparently true premises and an apparently false conclusion. At least one of these appearances is misleading, for the conclusion cannot be both true and false.

The argument is valid by the standards of orthodox modern logic. It is also valid by the standards of Stoic logic. Two logical principles are at stake. One is MP; it was the first indemonstrable (primitive) form of argument in Stoic logic: "If the first, then the second; but the first; therefore the second." The other is the "Cut" principle that valid arguments can be chained together: thus the valid argument from "1 is few" and "If 1 is few then 2 are few" to "2 are few" can be chained together with the valid argument from "2 are few" and "If 2 are few then 3 are few" to "3 are few," giving a valid argument from "1 is few," "If 1 is few then 2 are few" and "If 2 are few then 3 are few" to "3 are few." The relevant form of Cut was the third Stoic rule for the analysis of complex arguments: "If from two propositions a third is deduced and there are propositions from which one of the premises may be deduced, then the other premise together with these propositions will yield the conclusion."

On Stoic terms, the argument is valid, its first premise true and its conclusion false. Thus not all the conditional premises are true. By the Stoic principle of bivalence, at least one of them is false, despite appearances. At this point there is a complication. The truth and falsity conditions of conditionals were the subject of a fierce controversy that went back to Diodorus and his contemporary, Philo, and was taken up by the Stoics. Philo treated the

conditional "If *P* then *Q*" as a truth-function of its components equivalent to "Not: *P* and not Q." In contrast, Diodorus held "If *P* then *Q*" to be at least as strong as "Not ever: *P* and not Q." Chrysippus went still further; for him, a conditional is true if and only if its antecedent is incompatible with the negation of its consequent. Thus "If *P* then *Q*" is equivalent to "Not possible: *P* and not Q." In modern terms, Philo's conditional is material implication, Chrysippus's is strict implication. Later Stoics tended to follow Chrysippus.

In the sorites argument, some conditional premise "If *i* are few then *i* + 1 are few" is supposed to be false. If the conditional is Chrysippan, it is false if and only if "*i* are few" is compatible with "*i* + 1 are not few." However, this conclusion looks banal; who thought them incompatible? Chrysippus might cheerfully allow that all the conditional premises, so taken, are false. To know the falsity of such a conditional is not to identify a cut-off point; it is merely to know that a certain point is not debarred from being the cut-off. For some modern philosophers, sorites puzzles arise because vague concepts are subject to tolerance principles which do rule out the possibility of cut-off points and "*i* are few" does threaten to be incompatible with "*i* + 1 are not few," making the Chrysippan conditional "If *i* are few then *i* + 1 are few" true.[3] But the Stoics did not take that view, and may not have regarded the argument with Chrysippan conditional premises as genuinely challenging.

The most challenging form of the sorites argument uses the Philonian conditional, for it is the weakest connective to obey MP, which is to say that it is the weakest conditional. If the conditional premises are true on any reading, they are true on this one. Since it was not the standard reading of the conditional, the Stoics had to formulate the premises explicitly as negated conjunctions to confront the argument in its most telling form. Just that was done in standard Stoic accounts.

Once the explicit conditional has been eliminated, MP can no longer be used, but Stoic logic still obliges. The sorites argument with negated conjunctions is valid, its first premise is true and its conclusion false. Thus some premise of the form "Not: *i* are few and *i* + 1 are not few" is false. By the falsity condition for the Philonian conditional, *i* are few and *i* + 1 are not few. Thus *i* is a sharp cut-off point for fewness. Since one cannot identify such a point, one is in no position to deny any of the premises; one can only suspend judgment. The challenge "Which premise is false?" is unfair, for one may be unable to find out even though one knows that at least one premise is false.

What is gained by presenting the sorites as an argument with premises and conclusion? Its logical structure was never the point at issue, for the argument is formally valid according to those whom it threatens, the Stoics, who used arguments with that structure themselves. As for the Skeptics, they could suspend judgment on its logical status; it was enough for their purposes that their opponents took such arguments to be valid. The logical structure provides a convenient way of laying out the problem, but so far nothing more.

It is tempting to argue for a dialectical structure behind the logical façade. One point is that the use of conditionals in the sorites argument is a distraction, since the sorites interrogation shows that one can set the puzzle going in a language whose only resources are "Yes," "No," and simple sentences (without logical connectives such as "if," "and," and "not") in the interrogative mood. Moreover, the argument has been persuasive so far not because its premises commanded assent, but because they forbade dissent. The problem was not that one could say "Not: *i* are few and *i* + 1 are not few," but that one could not say "*i* are few and *i* + 1 are not few." One is not presumed to believe the premises of the sorites argument. The point of the questions is to force one to take up attitudes for or against the individual propositions, for any pattern of such attitudes leads one into trouble. Since the premises of

the sorites argument seem compelling only when one is interrogated on their components, the question form takes primacy.

The situation is transformed if the premises of the sorites argument can be given positive support. If they can, the argument form takes primacy: the question form leaves too much unsaid. What is more, Chrysippan silence is no longer an adequate response, for it does not undermine the positive support for the premises. Awareness of the need to provide that support is shown by Galen (AD *c*.129–*c*.199): "I know of nothing worse and more absurd than that the being and non-being of a heap is determined by a grain of corn." Chrysippus could not have suspended judgment on the general claim that one grain does not make the difference between a heap and a non-heap. He must deny it, for it contradicts the existence even of an unknown cut-off point. For him, the addition of one grain can turn a non-heap into a heap.

Galen's interest in sorites puzzles was connected with a long-running dispute between Empiricist and Dogmatic (one might say "Rationalist") Doctors. The Empiricist Doctors based their medical knowledge on inductive inferences, holding it to be reliable only if derived from sufficiently many observations; their opponents applied sorites reasoning against the notion "sufficiently many." The Empiricist Doctors replied that the argument proved too much; if it destroyed the notion of sufficiently many observations, it would by parity of reasoning destroy much of the common sense on which all must rely. They gave the examples of a mountain, strong love, a row, a strong wind, a city, a wave, the open sea, a flock of sheep and a herd of cattle, the nation and the crowd, boyhood and adolescence, the seasons: none would exist if sorites reasoning were to be trusted. The Empiricist Doctors could reasonably claim to know that sorites arguments were unsound, without claiming to know exactly where the flaw lay. Even Chrysippus could not say which premise in negated conjunction form was false.

It was known that for every sorites series which proceeded by adding (as Eubulides's original series seem to have done), a reverse sorites series proceeded by subtracting. Thus examples tend to come in pairs of opposites: rich and poor, famous and obscure, many and few, great and small, long and short, broad and narrow. The awareness of reversibility no doubt helped to check the tendency to think of a sorites puzzle as showing its conclusion to be strange but true, for the conclusion of one sorites argument contradicts the first premise of the reverse argument.

There are also signs of a rather different Empiricist point. The sorites questioner is compared to someone who asks a shoemaker what last will shoe everyone: the question has no answer, for different feet require different lasts. The idea may be that the required number of observations depends on the circumstances of the particular case. There is no general answer to the question, "Are 50 observations enough?" The point has been repeated by modern philosophers, and is correct as far as it goes, but that is not very far; for the questions can be asked about a particular case, and the Empiricist still cannot plausibly claim to know all the answers. The same goes for heaps. Fifty grains may make a heap in one arrangement and not in another; but in any particular process of piling up grains one by one there will be a point at which the right answer to the question "Is this a heap?" is unknown.

As logic declined in later antiquity, so did interest in sorites puzzles. They formed no part of the medieval logic curriculum, perhaps because of their absence from the works of Aristotle. Their revival had to wait for what is usually seen as the corruption of logic in the Renaissance. Lorenzo Valla (1407–1457) was one of the chief instigators of a shift from the formal rigor of scholastic logic to the more literary pursuit of humanist dialectic, and his

preference for Cicero over Aristotle led him to Cicero's account of sorites arguments. Valla rejected them, on the grounds that even one grain makes some difference, in his original treatment of multiple syllogisms. Unfortunately, later writers did not develop his suggestions. Sorites puzzles were known as curiosities. Logic textbooks used the term "sorites" for all syllogisms with more than two premises. Stoic arguments do not count as multiple syllogisms, for they turn on the logic of propositional connectives such as "if" rather than quantifiers such as "all" and "some." The analogy with the original sense of "sorites" lies in the repetition of steps, but the textbooks did not associate sorites syllogisms with paradoxes in any way.

Leibniz knew of the Heap and the Bald Man. For him, unlike the Stoics, the ideas of heap and baldness exemplify indeterminacy: their limits, like those of color ideas, have not been fixed. A borderline case is a matter of opinion, different opinions being equally good.[4] In one respect, Leibniz shared the view which has dominated twentieth-century discussions: that the ignorance associated with vagueness is not a matter merely of not knowing where the cut-off lies, but of there being nothing to know.

2 Recent Approaches

In the first half of the twentieth century sorites paradoxes aroused little interest. An exception is Russell's 1923 article. He takes the paradoxes seriously, and responds as follows:

> [if a hairy man goes bald] it is argued [that] there must have been one hair the loss of which converted him into a bald man. This, of course, is absurd …. there are men of whom it is not true to say they must either be bald or not bald. The law of excluded middle is true when precise symbols are employed, but it is not true when symbols are vague. (pp. 85–86)[5]

More recently, sorites puzzles have been discussed in the form of apparently sound arguments with apparently false conclusions, and philosophers such as Dummett and Wright have advanced grounds for the premises. As noted earlier, this renders Chrysippan silence by itself an insufficient response: the grounds for the premises must be rebutted.

The recent tradition has largely ignored the epistemic view of vagueness, assuming, by contrast, that borderline cases, initially characterized as those where we do not know what to say in answer to a sorites question, are cases in which there is no fact of the matter to be known. Such a view cannot avoid commitment to indeterminacy at the semantic level: for some sentences there is no fact of the matter whether they are true, and for some predicates and some objects, no fact of the matter whether the former apply, or fail to apply, to the latter. So we will refer to the "no fact of the matter" conception of vagueness as the "semantic" conception. (The conception may be coupled with different views about the ontological status of the absence of determinate fact. On one view, once the semantics have been properly formulated, there is nothing more to be said; on another view, the semantic indeterminacy reflects some real indeterminacy in the non-linguistic world itself.)

Developing the semantic conception of vagueness requires abandoning classical semantics or logic. In the context of the rise of formal methods dating from the middle of the last century, it is not surprising that a number of non-classical systems have been designed to accommodate vagueness, semantically conceived; and we will shortly survey some of these theories.

First, however, we present three schemata of sorites-paradoxical arguments (§2.1). Second, we examine arguments which have been given for the truth of their premises or the falsehood of their conclusions (§2.2). Third, we review some formal treatments of vagueness (§2.3); and fourth, we remind the reader of the attractions of the epistemic view of vagueness, but also indicate a potential source of difficulty for the view (§2.4).

2.1 Three Forms of Paradoxical Arguments

With no sand you cannot make a heap of sand; and if you have just one more grain, you cannot make a heap from what is not a heap. Hence there are no heaps of sand. The argument could be schematized so as to fit the standard pattern of argument by mathematical induction, with quantifiers ranging over the natural numbers and "x'" abbreviating "$x + 1$":

$$\text{(A)}\quad \frac{\varphi(0) \quad \forall x(\varphi x \rightarrow \varphi x') \quad \text{(QP)}}{\forall x \varphi x}$$

We refer to the quantified premise as QP. This schema yields the heap paradox itself if we replace "φx" by "a collection of x grains cannot make a heap no matter how it is arranged."

Imagine a painted wall hundreds of yards or hundreds of miles long. The left-hand region is clearly painted red, but there is a subtle gradation of shades, and the right-hand region is clearly yellow. The strip is covered by a small double window which exposes only a small section of the wall at any one time. It is moved progressively rightwards, in such a way that at each move after the initial position the left-hand segment of the window exposes just the area that was in the previous position exposed by the right-hand segment. The window is so small relative to the strip that in no position can you tell any difference in color between what the two segments expose. After each move, you are asked to say whether what you see in the right-hand segment of the window is red. You must certainly answer "Yes" at first. At each subsequent move you can tell no difference between a region you have already called red and the one for which the new question arises. It seems that you must after every move call the next region red, and thus, absurdly, find yourself calling a clearly yellow region red.

This form of sorites can also be molded to schema (A). One could stipulate that the successive positions of the right-hand segment of the window are numbered upwards from 0, and replace "φ" by "… numbers a red region." Equally, and with no difference of substance, we can regard the numerals as simply naming the successive regions, and replace "φ" simply by "red."

One does not need to argue by mathematical induction to have sorites paradoxes. We could, as we saw in §1 (p. 738) write out, say, 10,000 singular conditionals instead of QP, and use *modus ponens* to derive the absurd result that a collection of 10,000 grains, however arranged, cannot make a heap. The schema is:

$$\text{(B)}\quad \frac{\varphi(0) \quad \varphi(0) \rightarrow \varphi(1) \quad \ldots \quad \varphi(9{,}999) \rightarrow \varphi(10{,}000)}{\varphi(10{,}000)}$$

We do not need to use either conditionals or quantifiers in the premises, for example:[6]

$$\begin{array}{ll} (\mathrm{C}) & \varphi(0) \\ & \neg\varphi(10{,}000) \\ \hline & \exists x(\varphi x \,\&\, \neg\varphi x') \quad (\mathrm{QC}) \end{array}$$

An instance might be: A man with 0 hairs on his head is bald, and a man with 10,000 hairs on his head is not; therefore there is some number of hairs such that a man with that number of hairs is bald and a man with one more hair is not. The supposed absurdity is the derivation of the existence of a sharp boundary, represented by the quantified conclusion (QC).

In all cases, the arguments appear to be sound, yet they have what appear to be false conclusions.

The availability of these different formulations, and others we have not mentioned, puts constraints on what sort of solution is acceptable. A reasonable initial hope is that there is a single correct approach to all forms of sorites. If this hope can be fulfilled, it would be no good merely to argue, with schema (A) in mind, that the principle of mathematical induction does not hold for vague predicates, for this would not touch sorites arguments exemplifying schemata (B) and (C).[7] Equally, it would not do to suppose that one need only point out that by adopting an intuitionistic logic, one can deny the QP of (A) without thereby being committed to the seemingly unacceptable classical equivalent of its negation, $\exists x(\varphi x \,\&\, \neg\, \varphi x')$ (cf. Putnam, 1983; Read and Wright, 1985; Putnam, 1985); for paradoxical arguments modeled on the other schemata are not touched by this point.

2.2 Arguments for the Premises or against the Conclusions

We will consider two kinds of argument: (a) from the nature of observation; and (b) from the nature of vagueness.

2.2.1 Observational Predicates

Perhaps one can always apply "red" or at least "looks red" correctly, under suitable circumstances, just by looking. If so, then if under such circumstances two things look the same, and the predicate is true of one, it is true of the other. With "red" or "looks red" inserted as before into the replacement for "φ," this line of thought justifies the conditional premises in schema (B), the QP in (A), and the negation of the conclusion of (C) (that is, not-QC). The conditionals would be justified by the fact that, given that adjacent regions on the wall look the same, the truth of an antecedent of the form "x is (or looks) red" would be enough for the truth of a consequent corresponding to "x' is (or looks) red." The QP of (A) would be justified by the fact that, given the same feature of the wall, "red" (or "looks red") must apply to both or neither of adjacent regions; and this would also establish the falsehood of the QC of (C).

Since the truth-values of the atomic premises and conclusions are uncontroversial, this would take care of all that is controversial in showing that the arguments have true premises and false conclusions. If one takes this result at face value one must regard classical logic, and any other logic upon which arguments molded on any of the schemata (A)–(C) are valid, as incorrect.

"Red" (or "looks red") is meant to be an example of a more general phenomenon: the "observationality" of certain predicates. Dummett suggests that an observational predicate is one "whose application is determined by mere observation" (1975, p. 261), and "can be

decided merely by the application of our sense-organs" (p. 265). Dummett takes it to be a consequence of the observationality of a predicate like "red" that it is "governed by the principle that, if I cannot discern any difference between the colour of *a* and the colour of *b*, and I have characterized *a* as red, then I am bound to accept a characterization of *b* as red" (p. 264). The justification is that a predicate could not be applied simply on the basis of how things look, and so could not be observational, if an indistinguishable difference could determine the predicate's differential applicability.

To use this thought to ensure paradoxical truth-values for the elements of sorites arguments, we could define the observationality of φ as follows:

> if α satisfies φ and, under normal circumstances for observation for φ, β is indistinguishable from α, then β satisfies φ.

If adjacent members of a sorites series count as indistinguishable, this ensures that observational predicates are paradoxical. But it is questionable whether there are any non-trivial observational predicates, thus defined. Consider "looks red" as a strong candidate. Suppose that α and β are indistinguishable under, say, standard lighting conditions, and so on. Then one might argue: their indistinguishability entails that they look the same; so if α looks red, so does β; so the condition for "looks red" being observational is met. So there is at least one observational predicate.

The notion of indistinguishability is crucial to this argument, but it has not been adequately explained. To show the complexity, consider a case in which regions on the wall are exposed to a subject in a random order (the "random sorites"), and he is asked to say of each region whether it looks red. We then score the result by writing a "Y" beneath a region if the subject said "Yes" with respect to that region, an "N" if he said "No," and a "0" otherwise. What would the score of a perfectly rational subject look like, assuming conditions of observation to be normal for "looks red," and assuming that his eyesight and mastery of the predicate are both perfect? There will certainly be a "Y" beneath at least one of the left-hand regions, and an "N" beneath at least one of the right-hand regions; there will be no "0" to the left of a Y, and no "N" to the left of a "0." But there will certainly be a Y-region, call it α, which has a "0" or "N" adjacent to it and to its right, that is, at α'.

Given the assumptions about the rationality of the subject and the perfect conditions for observation, one might conclude that he has not made a mistake about how things looked to him, even though distinct rational subjects, and the same subject on different occasions, will draw the line in different places. This means that there is no mistake in supposing that α satisfies "looks red" and α' does not. We could infer that adjacent regions of the wall are not, after all, indistinguishable, for a rational subject, without error, made a distinction between them. In that case, we will be hard put to find sorites series with indistinguishable members, and so will be hard put to use the notion of observationality to justify the premises of arguments of types (A) and (B).

Alternatively, we might insist that the adjacent regions of the wall are indistinguishable (for example, because they look the same when co-presented), but then we must say that "looks red" does not satisfy the envisaged condition for being an observational predicate.

Either way, observationality has no impact on the sorites paradoxes.[8] This is what one would have predicted on the supposition that all sorites paradoxes should be accounted for in the same way, for there are many sorites-paradoxical predicates which have little claim to observationality, such as "child," "dog," "know," and "few." Even for those for which observationality is

at issue, like "red" or even "heap," we can generate a sorites series in which the adjacent members are discriminable. Intuitively, a patch which differs only just discriminably from a red patch should count as red; and, as we saw, the paradox of the heap is quite impervious to the assumption that the various collections differ in size, in a potentially knowable way.

2.2.2 The Nature of Vagueness

The semantic conception of vagueness holds that it is of the nature of a vague predicate to draw no sharp boundary between the things to which it applies and those to which it does not. Arguably, this could be expressed precisely as the QP of schema (A), as the conditionals of schema (B), and as a denial of the QC of schema (C). So we have an argument for the paradoxical distributions based on a claim about the nature of vagueness.

Resisting this argument requires, from the perspective of the semantic conception, adjustments to classical logic or classical semantics or both. Perhaps one can persuade oneself that an adequate account of the nature of vagueness should not entail QP. We don't believe a tadpole can turn into a frog in a millisecond, but perhaps we can refrain from holding that it is in general true that anything which is a tadpole at a time is also a tadpole a millisecond later: we see too clearly where that belief would lead. So perhaps there is no immediate argument from the nature of vagueness to the truth-value distributions required to make arguments of type (A) paradoxical.[9] But there is a less immediate argument. Suppose we are happy to resist QP, that is, to hold that it is not true. Then the natural thing to say is that it is false; but this commits us to the truth of its negation, which is classically equivalent to QC: $\exists x(\varphi x \,\&\, \neg\, \varphi x')$. So we sever a direct connection between the nature of vagueness and the paradoxical truth-value distributions, only to force into view a less direct connection. Perhaps we could hold that QP is not true, yet not that it is false; perhaps this would be enough to avoid the unwelcome QC and also avoid the paradox-inducing QP. However, we would need an account of an appropriate non-classical semantics, and a justification for the specific treatment of QP.[10] The general moral, then, is that a semantic view of vagueness requires non-classical semantics and/or non-classical logic.

2.3 Alternative Logics and Semantics

On the semantic view, it would be desirable to be able to say that QP is not true without being forced to say that QC is true; and desirable to be able to say that at least one conditional from schema (B) fails, without having to say that one of them is false. (The envisaged conditionals are material, so a false one has a true antecedent and a false consequent.)

To do this, one would have to abandon bivalence. There are two ways to proceed. We can use a bivalent metalanguage to describe a non-bivalent object language. Well-known examples are given in §2.3.1 and §2.3.2. Or we can use a non-bivalent metalanguage to describe our non-bivalent object language. An example is sketched in §2.3.3. On the first alternative, we can keep to classical conceptions of sets and models: vagueness is then tamed, described in essentially non-vague terms. On the second alternative, vagueness is never eliminated, never sharply described.

2.3.1 Supervaluations

If an expression is vague, we believe it could, in theory, be replaced by a more precise one. Thus we could replace the vague "child" by the more precise "minor" (meaning person who has not yet reached the day of his or her eighteenth birthday). Not every replacement would

be acceptable: if "minor*" is defined as a person who has not yet reached the day of his or her fifth birthday, the word is nowhere near "child," since it would be clear that some children are not minors*. Equally, the expression "minor′," defined as a person who had not yet reached his or her thirtieth birthday, would be unacceptable, since it would be clear that some minors′ are not children.

An important line of thought about vagueness is that this notion of an acceptable way of making an expression more precise can be used to specify the meaning of a vague expression: its meaning will be given in terms of all the ways in which it could acceptably be made precise, where an acceptable way is one not precluded by the meaning it has. An acceptable way does not make the extension of the term too narrow, like that of "minor*," or too wide, like that of "minor′." A sentence containing a vague expression is to count as true, if it is true however that expression is made precise, and to count as false if it is false however that expression is made precise. This makes room for cases in which the sentence is true on some ways of making it precise, false on others; and, hence, in which it is neither true nor false.

In a classical model, a (unary) predicate is associated with a set of entities from the domain. Using the idea of the previous paragraph, we could associate a vague unary predicate with a range of sets, the sets corresponding to the acceptable ways in which it could be made precise.

This approach has been developed with formal elegance in a kind of model theory called supervaluation theory.[11] In addition to starting from an appealing conception of vagueness, it offers two further charms: it promises a precise description of a vague language; and it promises to preserve classical logic.

Let us think of a model, M, for a language, L, as an ordered pair $\langle D, F \rangle$ of a domain, D, and a family F of valuation functions, each mapping each unary predicate of L on to a subset of D and each name of L on to a member of D. (For simplicity, we will assume that all the non-logical symbols of L are either unary predicates or names.) Suppose that for some predicate, φ, of L, every member f of F meets these conditions:

if x is in D and is definitely a satisfier of φ then x is in $f(\varphi)$.
if x is in D and is definitely not a satisfier of φ then x is not in $f(\varphi)$.

Then we shall say that M is *appropriate* for φ. A model is appropriate for a language iff it is appropriate for all its expressions.[12] An appropriate model is *maximal* iff any addition to its family of valuations would render it inappropriate. Maximal models are supposed to represent the semantics of languages with vague predicates.

Within any model, truth-relative-to-a-member-f-of-F (abbreviation: truth$_f$) is defined in the usual classical way:

⌜φα⌝ is true$_f$ iff f(α) is in f(φ);
⌜A & B⌝ is true$_f$ iff f ⌜A⌝ is true$_f$ and ⌜B⌝ is true$_f$;
... and so on.
⌜A⌝ is false$_f$ iff ⌜A⌝ is not true$_f$.

Truth and falsity (relative to a model) are defined by generalizing over valuation-relativized truth: ⌜A⌝ is true in M iff ⌜A⌝ is true$_f$ for every f in F; ⌜A⌝ is false in M iff ⌜A⌝ is false$_f$ for every f in F. Truth and falsity are thus "supervaluations" relative to the basic bivalent valuations in F. This allows for borderline cases to induce a failure of bivalence in maximal

models. Suppose that a member α of D is not definitely a satisfier of φ and also not definitely not a satisfier of φ. And suppose that $M = \langle D, F \rangle$ is maximal. This means that there is a member of F, say f_1, such that $f_1(\varphi)$ contains α, and a member of F, say f_2, such that $f_2(\varphi)$ does not contain α. So "φα" is true_{f1} and false_{f2}. So "φα" is neither true nor false in this model.

Although supervaluation theory treats the object language as non-bivalent, the Law of Excluded Middle is preserved, that is, the schema $\ulcorner A \vee \neg A \urcorner$ is valid. Consider the instance of it with φ and α as in the previous paragraph. It will be true_f for every f, since for every valuation either $f(\alpha)$ is in $f(\varphi)$ or it is not, and in the first case the disjunction will be true_f in virtue of the truth_f of its first disjunct, and in the second case in virtue of the truth_f of its second. In supervaluation theory, a disjunction can be true without either disjunct being true.

Validity can be defined as usual: an argument is valid iff every model in which all the premises are true is one in which the conclusion is true. The class of valid arguments thus defined is identical with the class of classically valid arguments.[13] Hence arguments following schemas (A), (B), and (C) are all supervaluationally valid; but sorites instances of schemas (A) and (B) are unsound. We will first show how this is so, and then consider what the supervaluational theory has to say about instances of (C).

By the definition of appropriateness, the valuations in an appropriate maximal model will place intuitively definite cases in the extension of any vague predicate, and exclude intuitively definite non-cases from its extension. For "heap," each f will associate the predicate with a set of collections (of grains of sand, say), where one collection is the smallest, and the set also contains all larger ones.[14] In other words, every f will associate the predicate with a sharp threshold, though, because of the predicate's vagueness, the different valuations will associate it with different thresholds. For any f in F, one conditional is false_f: the conditional which, for the line drawn by f, has an antecedent referring to an object on one side of the line and a consequent referring to an object on the other. Hence at least one conditional (typically, several conditionals) in the premises of schema (B) will fail to be true.

Similar facts ensure that QP in schema (A) is not merely not true, but false. The singular instances of QP are the conditionals which are the premises of (B). Since every f in F falsifies one of these conditionals, every f falsifies QP.

Many objections have been leveled at supervaluational theories. We list two.

The Truth of QC

The supervaluational theory has it that instances of (C) are sound, and thus have a true conclusion, QC: $\exists x(\varphi x \,\&\, \neg\, \varphi x')$. If we find this hard to swallow, we are asked to remember that, in the presence of vagueness, quantifiers do not work in the normal classical way. In particular, using "satisfies" on the lines of "true," there is no sound inference from the truth of QC in the model to the conclusion that there is something in the domain of the model which satisfies "$\varphi x \,\&\, \neg\, \varphi x'$." The supervaluation theorist could claim that only the existence of such an object would amount to the existence of a sharp cut-off; so the truth of QC does not entail a sharp cut-off; so one is confused if one thinks that one wants to deny QC.

By itself, this is unlikely to be very persuasive. It seems plain that one can use our ordinary language to express absence of a cut-off, without having to ascend into the metalanguage, as this response supposes.

However, it is plausible that an object-language account of what it is for a predicate to be vague involves, at least in the first instance, the use of some expression representing definiteness, or, equivalently, vagueness. One wants to say, for example, that "red" is vague because there are things which are neither definitely red nor definitely not red; or that it is

vague because there are things concerning which there is no fact of the matter whether they are red or not. More formally, we might say that any vague language could be expected to be capable of expressing its vagueness, perhaps by a sentence operator, "Def," expressing definiteness.[15]

The supervaluation theorist would do well to claim that something like this operator must be used in saying what we want to say about vague predicates and cut-offs: what we really want to say is that there is no *definite* cut-off point. This is properly expressed not by denying QC, but rather by denying "$\exists x\mathrm{Def}(\varphi x \,\&\, \neg\, \varphi x')$." This means extending supervaluation theory to deal with "Def."

From a supervaluationist's point of view, "Def" should resemble an object-language expression of the notion of truth (by supervaluational lights).[16] Thus "Def A" should be true-on-a-valuation iff A is true (that is, true-on-every-valuation). This makes "Def" in some ways like "$\Box$" ("$\Box$A" is true-at-a-world iff "A" is true-at-every-world). In particular, taking the analogy in the most straightforward way, "Def" would eliminate vagueness: applied to any sentence, the result is one which is true on all or on no valuations.

However, there are two important points of disanalogy between "Def" and "$\Box$." First, if the supervaluationist retains the definition of validity as, in effect, necessary preservation of truth, he must agree that once "Def" is added to the object language, certain classical forms of reasoning (conditional proof, *reductio ad absurdum*, and *or*-elimination) cannot be allowed to be valid.[17] Second, if there is higher-order vagueness, then vagueness should not be eliminated by "Def."

Problems with Higher-Order Vagueness

A supervaluational model is intended to divide the sentences of the language into three sets: the truths, the falsehoods, and the remainder. There will be adjacent members, α and α', of a sorites series such that α is the last truth in the series and α' the first non-truth. But many find this unacceptable: they claim that there is no more a sharp boundary between the truths and the borderline cases than there is between the truths and the falsehoods.

The notion of a supervaluation is itself vague: it is defined in terms of an acceptable model, which is in turn defined in terms of what definitely falls under a predicate and what definitely does not. These notions admit of borderline cases: the reasons for saying that, for example, there is no fact of the matter whether certain color patches are red support with no less strength the conclusion that there is no fact of the matter whether certain patches are *definitely* red. Equally, the reasons for thinking that there is no last region of the wall which counts as red support with no less strength the conclusion that there is no last region on the wall which is definitely red. If the conclusion is correct, there will be a valuation such that there is no fact of the matter whether or not it is appropriate.

One ensuing problem is that in an object language in which there is higher-order vagueness, "Def A" is neither true nor false for some A; so the truth-conditions for "Def" cannot be given in the simple way envisaged earlier ("Def A" is true-on-a-valuation iff A is true). One response draws inspiration from possible-worlds semantics for modality: one would need to introduce an analogue of the accessibility relation, holding between valuations, and this would need to be reflexive (to validate "Def A → A") but not transitive (in order not to validate "Def A → DefDef A").

A deeper problem is that the supervaluationist's notion of truth as truth-on-all-appropriate-valuations must itself be vague. This means that a potential charm of supervaluation theory is lost: it cannot, after all, give a precise description of a vague language. It

also raises the question whether the supervaluational concept of truth is correct. Many believe that truth must satisfy Tarski's disquotation schema:

(T) True (A) iff A.

But then, by reasoning which the supervaluationist endorses, it follows that if A is not true it is false.[18] The supervaluationist category of neither-true-nor-false sentences would vanish, and with it the basic idea of supervaluation theory.[19] So the supervaluationist must deny that his concept of truth is disquotational, and this raises the question whether it is, properly speaking, a concept of *truth* at all.

2.3.2 Degrees of Truth

Not all borderline cases for a vague predicate are equal (or so it may reasonably seem). Two patches α and β might both be borderline cases of red, yet one redder than the other: one has a greater degree of redness. So one might be tempted by the following progression: it is closer to the truth to say that α is red than to say that β is red; so truer to say that α is red than to say that β is red; so "α is red" has a greater degree of truth than "β is red."

It is certainly an important feature of many vague predicates, and perhaps of all those which are sorites-susceptible, that the relevant cases are subject to an underlying comparative relation. However, this cannot be taken as any kind of knockdown argument for the existence of degrees of truth, for we have yet to distinguish between the relatively innocuous suggestion that some sentences are nearer to stating the (absolute) truth than others, and the controversial suggestion that some sentences are truer than others. We might make some progress towards the more exotic suggestion by reflecting on the following lines: truth is what one should aim at in belief, but for borderline cases the best you can aim at is degrees of belief; so we must be able to make sense of a notion of degrees of truth.

Whatever the philosophical motivation, semantics based upon degrees of truth have been claimed to dispel sorites paradoxes.[20] In the models $\mu = \langle \delta, \Phi \rangle$ which we now consider, δ is a domain of individuals, and Φ assigns to every name in the language a member of δ, and to each predicate, φ, a J-set, $\Phi(\varphi)$. A J-set is a mapping of members of δ into the real numbers in the closed interval [0,1]. Intuitively, the idea is that the number which is the value of the mapping for the member of δ which is the argument – the "J-value of $\Phi(\varphi)$ for δ" – represents the degree of truth, relative to assignment Φ, associated with affirming φ of that member. The value 1 represents an object to which the predicate definitely applies, the value 0 represents an object to which the predicate definitely does not apply, and the intermediate values represent intermediate cases. Writing $[P]_\Phi = n$ to express the fact that the sentence P is assigned the degree of truth n by function Φ, atomic sentences are assigned degrees of truth by rules like: $[\varphi\alpha]_\Phi$ is the J-value of $\Phi(\varphi)$ for argument $\Phi(\alpha)$. Appropriate models assign J-sets which reflect actual usage, that is, lead to assignments of degrees of truth to atomic sentences which both preserve our intuitive orderings (such as our ordering of how red the patches on the wall are) and conform to our intuitions about definite truths (assigned 1) and definite falsehoods (assigned 0).

The theories we discuss treat the logical constants as degree-functional: that is, the degree of truth of a complex is a function of the degrees of truth of the constituents. There are various possible functions. One standard approach is to stipulate as follows:

$$\begin{aligned}
&\left[\neg \mathrm{P}\right] = 1 - \left[\mathrm{P}\right] \\
&\left[\mathrm{P} \,\&\, \mathrm{Q}\right] = \min\left\{\left[\mathrm{P}\right], \left[\mathrm{Q}\right]\right\} \\
&\left[\mathrm{P} \vee \mathrm{Q}\right] = \min\left\{\left[\mathrm{P}\right], \left[\mathrm{Q}\right]\right\} \\
&\left[\mathrm{P} \rightarrow \mathrm{Q}\right] = 1, \text{if} \left[\mathrm{Q}\right] \geq \left[\mathrm{P}\right], = 1 - \left(\left[\mathrm{P}\right] - \left[\mathrm{Q}\right]\right) \text{otherwise.} \\
&\left[\forall v \mathrm{A}v\right] = \mathrm{glb}\left\{\left[\mathrm{A}^{a} / v\right] : \text{for all a}\right\} \\
&\left[\exists v \mathrm{A}v\right] = \mathrm{lub}\left\{\left[\mathrm{A}^{a} / v\right] : \text{for all a}\right\}^{21}
\end{aligned}$$

The functions give the classical results if the arguments are restricted to 1 and 0. The specific idea behind the equation for ¬ is that there is no difference between departing from definite truth (that is, value 1) and approaching definite falsehood, and that predicates have paradigm borderline cases (such as α for φ) for which we want [φα] = [¬φα]. The specific idea behind the equation for → is that a conditional should fall short of perfect truth to the extent that truth leaks away as between antecedent and consequent.

One standard account of validity within degree theory is based on a generalization of the notion of truth-preservation: a valid argument is one such that every model assigns a degree of truth to the conclusion no lower than that assigned to the lowest-valued premise. However, Edgington (1992) has argued that it would be better to give a different account, based on the idea that validity does not permit additional falsehood: a valid argument is one such that every model assigns a degree of falsehood to the conclusion which does not exceed the sum of the degrees of falsehood it assigns to the premises.

One impact of this view upon the sorites is that arguments of type (A) and type (B) will be seen as having at least one premise which is not entirely true. Thus the truth-values of the components of the conditionals of (B) progressively fall, and a conditional with a consequent less true than its antecedent will not be entirely true. For similar reasons, the QP of (A) is also not entirely true. Hence, on this approach, the impression that the sorites arguments are sound is, for these cases, supposedly dispelled by dispelling the appearance of (fully) true premises.

On the more common definition of degree-theoretic validity, generalizing from truth-preservation, none of the argument patterns we have discussed is valid. For example, type (B) is not, because *modus ponens* is, by this standard, not valid in degree theory. For suppose that $[P] = 0.9$, and $[Q] = 0.8$. Then $[P \rightarrow Q] = 0.9$. So an argument of the form:

$$P, P \rightarrow Q \therefore Q$$

has a lowest-valued premise of 0.9, and a conclusion valued 0.8.

Whichever definition of degree-theoretic validity one uses, arguments of type (C) are invalid, for the premises will have degree 1 and the conclusion a degree of around 0.5. So the apparent soundness of some versions of the sorites arguments is supposedly dispelled, through dispelling the appearance of validity.

However, the degree-theoretic account is open to various objections. We mention three.

It Does Not Do Justice to QC

In connection with (C), natural assumptions about the existential quantifier in degree theory have the result that QC does not come out as false (=0) but only as midway true. A defense of degree theory would have to claim that what should really have degree 0 is not QC itself, but rather "$\exists x\mathrm{Def}(\varphi x \,\&\, \neg\, \varphi x')$." The required extension of the theory to an object language containing "Def" is not straightforward. If the sentence just quoted is to receive degree 1, it would seem that [DefA] = 0 iff [A] < l. This means that "Def" would eliminate vagueness, and thus would not allow for higher-order vagueness.

Its Logic is Unintuitive and Unmotivated

People have found the assignments of degrees to complexes unintuitive and unmotivated. For example, "P v $\neg\, P$" will be as true as "P & $\neg\, P$" when $[P] = 0.5$. Similar unintuitive results are obtained when the sentences stand in non-formal logical relations. Thus suppose Eve is definitely female but a borderline case for being an adult, so that [Eve is an adult] and [Eve is a woman] are both 0.5. Degree theory cannot distinguish between "Eve is a woman if and only if Eve is an adult" and "Eve is a woman if and only if Eve is not an adult," assigning both 1.[22]

Even certain sentences which may be critically involved in some versions of sorites arguments are treated by this degree theory in an unintuitive way. Thus a classical equivalent of QP, "$(\forall x)\neg\, (\varphi x \,\&\, \neg\, \varphi x')$," will be assigned a value of about 0.5, whereas intuitively it ought to come out as nearly true.

It Does Not Do Justice to Higher-Order Vagueness

The degree-theoretic property of having degree of truth 1 is supposed to correspond to some intuitive property (or else the formal semantics cannot connect with the informal judgments of truth and validity which underlie sorites paradoxes). Perhaps it is that of being true, or of being definitely true, or of being completely, definitely, and unimpugnably true. Whatever property we choose, we have something unsatisfactory: since having degree 1 is a sharp property, the corresponding truth-related intuitive property must be sharp too. In any sorites series there is a last sentence having truth to degree 1; hence there is also a last sentence with the relevant truth-related intuitive property. But it is natural to suppose that there is no such sharp cut-off: whatever property we consider, true to a certain degree, absolutely true, definitely true, or completely, definitely, and unimpugnably true, there are no adjacent members of the sorites series one of which lacks, while the other possesses, this property.[23]

2.3.3 A Non-bivalent Metalanguage

It is hard to see how there could be a more accurate account of what you should aim to do, in using a vague word like "red," than that, if you want to keep to the truth, you should apply it just to red things. This suggests that we should look for a semantic theory which unashamedly uses vague vocabulary in the metalanguage. In particular, we wish to mention a theory recently proposed by Michael Tye (1994).[24]

His proposal is based on three semantic values: true, false, and indefinite. The connectives are assigned values as follows: a negation is true iff its component is false, false iff its component is true, otherwise indefinite; a conjunction is true iff both components are, false iff one component is, otherwise indefinite. Universal quantifications are true iff all their instances are, false iff an instance is, otherwise indefinite. So the premises of sorites arguments of schemata (A) and (B) are indefinite, and not true. These argument-patterns are not sound.

Merely having three values is of no help in itself, once higher-order vagueness is admitted, as we saw in the case of supervaluation theory. For example, suppose that in a sorites series of sentences there is a last which is assigned truth. It makes no difference whether the next sentence is assigned falsehood or some other value: either way, we have a boundary where, according to many intuitions, no boundary should be. Tye deals with this by saying that there is no fact of the matter whether all sentences have one of the three truth-values. Hence the claim that, for example, there is a last true sentence in a sorites series is vague, and not true. Once again, we would not have the making of soundness in sorites arguments. If this strategy works at all, it should work as well with just two truth-values, distributed in a similarly vague way.

To obtain truth-conditions for atomic sentences, Tye associates a predicate with a vague set, a set concerning which there are things for which it is vague whether they are members of it. Thus the approach involves recognizing vague extra-linguistic entities, which is controversial.

A problematic feature of his account is that indefiniteness in a component of a complex sentence can render the whole sentence indefinite regardless of its form. Thus "if *P* then *P*" is indefinite if *P* is, yet many find it compelling to regard it as true however things are with *P*. To take an example closer to the sorites, consider the reverse of QP:

RQP $\quad \forall x\left(\varphi x' \rightarrow \varphi x\right)$

Applied to the wall, this says that if a region further to the right (towards the yellow end) is red, so is the adjacent region to the left; so it is intuitively true. But on Tye's semantics, it is not true, since the truth of a quantification requires the truth of its instances, and some of these conditionals will have indefinite components, and thus be themselves indefinite. Intuitively, however, RQP is true, and corresponds to what some would see as a constitutive principle: anything redder than a red thing is red.

It is not clear which features of Tye's account are essential. As we have mentioned, the third value could not be essential to dealing with the sorites (even if it is required for the best description of the use of language). It may be that the apparently implausible features can be removed, or shown not to be implausible after all. However, if higher-order vagueness is a genuine semantic phenomenon, there cannot be a precise semantic theory for a vague language. It is plausible to infer, as Tye in effect does, that in this case we must abandon a conception of semantics as making, for every sentence of a language, some definite pronouncement (such as that it is true, false, or indefinite): there will have to be sentences for which there is no definite fact of the matter concerning what the semantic theory says about them, and this must be distinguished from the semantic theory saying that there is no definite fact of the matter concerning their semantic value. Thus Tye's approach requires semantic theories to be, in a sense, incomplete.

2.4 The Epistemic View[25]

All these problems would be solved on the epistemic view, according to which there really are sharp cut-offs, though we cannot identify them. We cannot identify them because we do not have the required fineness of discrimination: we must allow for a margin of error in our cognitive mechanisms. If we know of a region of the wall that it is red, we could not know of its neighbor to the right that it is not red (even if in fact it is not red); for we could not reliably distinguish shades which differ so little.

Arguments of type (C) are seen as sound demonstrations of the epistemic view. Arguments of types (A) and (B) are valid on the view, and establish the falsehood of a premise by *reductio ad absurdum*. There is no problem about denying QP and accepting QC, and no problem about regarding one of the conditionals in type (B) arguments (though one does not know which) as having a true antecedent and false consequent. One can retain all the simplicity of classical logic and semantics.

The epistemic view thus has plenty to be said for it, yet it often evokes incredulity. Those who see any merit in the rule-following considerations advanced by Wittgenstein, and his denigration of the notion of imperceptible rails upon which the correct usage of language is supposed to run, are likely to be particularly hostile to the epistemic position. However, it is easier to identify generalized hostility than precisely formulated opposing arguments. In this section, we try to give voice to two connected objections.

There is No Evidence that Vague Concepts Induce Sharp Cut-Offs

The point of this objection is to claim that the main reasons in favor of the epistemic view are quite general, like arguments of type (C), whereas what would be needed would be a detailed examination of how specific vague concepts actually work. The suggestion is that nothing in the details of how they work would ground the view that they are associated with sharp cut-offs.

However, at least for some vague concepts, close examination reveals some surprisingly rich cut-off-determining principles. Consider that paradigm of vagueness, "heap."[26] A heap of φs must be heaped up, and this involves at least one φ stably above another, but not in virtue of glue-like attachment. Thus grains of sand spread out in such a way that no grain is on top of another cannot be a heap, however many grains there are. Moreover, gluing grains together in such a way that some grains are on top of others is not a way of making a heap. It seems to be of the essence of heaps that they are held together by gravity alone. The arrangement must also not be that of a stack: a dry stone wall is not a heap of stones. If we think of roundish things like grains of sand, it would appear that the smallest stable arrangement meeting these conditions requires four of them: three grains close together supporting a fourth on top. Here we arguably have a sharp cut-off. (The governing principles may well not determine the same cut-off for every shape of object.) An argument of this kind is at best suggestive, and it is always open to the conventional theorist to say that we were wrong in counting as vague some predicates for which such cut-off-determining principles can be discovered. The conventional theorist may be on safest ground with color predicates.

The Epistemic View Does Not Do Justice to the Fact that Meaning Supervenes on Use

"Meaning is use": that is, semantic facts concerning a language supervene on facts about the linguistic behavior of masters of that language. Can the epistemic view do justice to this?

By "meaning supervenes on use" one might mean just that if two communities used a language in just the same way, then every sentence of the language would have the same meaning in both communities. By this standard, the epistemic theory of vagueness can certainly claim that meaning supervenes on use.

A much more stringent demand is that a theorist should provide explicit details of how meaning supervenes on use. Since no theorist of any kind (whether or not vagueness is at issue) has given any such detailed account, the fact that the epistemic theorist has not should not count against the theory.

However, the supervenience doctrine can be used to generate demands of a strength intermediate between the extremely weak and extremely strong demands of the previous two paragraphs. For there may be *a priori* principles relating to conditions under which supervenience is possible.

One proposed principle of this kind is verificationism. On the verificationist view, meaning can supervene on use only through knowledge: a sentence cannot have a meaning such that it would be impossible for those who understand it to determine whether it is true or false. Since the epistemic view holds that it is impossible for us to know where the cut-offs come, the view is inconsistent with the verificationist constraint upon the supervenience relation. However, since verificationism, at least in the strong form envisaged, now has few supporters, it does not pose a serious threat to the epistemic view.

One does not have to be a verificationist to feel qualms about whether the epistemic view can do justice to the supervenience of meaning on use. For example, one might hold that if a predicate stands for a manifest property (one which under some conditions detectably obtains), then under optimal conditions for manifestation, if there is a fact of the matter whether or not it obtains, that fact is detectable. This is not full-blown verificationism, for it is consistent with there being properties which are not manifest, and consistent with it being impossible for us to detect even manifest properties, to the extent that it is impossible for us to bring about optimal conditions. Yet, it might be claimed, we can view a region of the colored wall under optimal conditions without being able to detect the presence or absence of redness; and the view then delivers that there is no fact concerning whether the region is or is not red.

Clearly the epistemic theorist will take issue with this line of thought, challenging, among other things, what is taken for granted in the notion of optimal condition for manifestation (cf. Williamson, 1994, pp. 180–184). Until these issues are clarified and resolved, some caution about the epistemic view is called for. However, we are not aware of any decisive refutation of it, and it would provide a breathtakingly simple solution to sorites paradoxes.[27]

Notes

1 The account of the sorites in antiquity draws heavily on Barnes (1982) and Burnyeat (1982). The most important ancient texts are translated in Long and Sedley (1987, vol. 1, pp. 221–225). For more on the history of the sorites see Williamson (1994).

2 The Stoics were not sure that any wise men had lived; they may also have held that only a wise man would know anything at all. Nevertheless, knowledge was what the Stoics aimed at. The argument in the text could still be made with "justified true belief' in place of "knowledge."

3 See below, §2.2.

4 Leibniz (1961, III v 9 and III vi 27); see Wiggins (1980, p. 124).

5 It is now customary to distinguish between the Law of Excluded Middle, which requires the validity of the schema "Either *A* or not *A*," and the principle of bivalence, of which one formulation is that every sentence is true or false. (In §2.3.1 below we describe a theory which preserves the Law while rejecting the Principle.) Russell is best understood as rejecting the latter.

6 For this formulation, see Rolf (1984, p. 220).

7 Cf. Dummett (1975, pp. 251–252). Kamp (1981, pp. 226–27) provides an independent reason for the same conclusion.

8 The reason offered here for this conclusion can be found in Wright (1987, p. 283, especially n. 13) and Williamson (1990, pp. 88–103).

9 In particular, it is reasonable to suggest that our intuitions about vagueness are properly expressed only using some expression for definiteness. Cf. Wright (1987; 1992).

10 We would, in particular, need an account which allowed that something not true can yield a falsehood, without itself being false.

11 This form of theory was systematically applied to vagueness by Dummett (1975), Fine (1975), Kamp (1975), and Lewis (1970). Fine traces the origins of the idea, as applied to vagueness, to Mehlberg (1958). The use of the expression "supervaluation" goes back to van Fraassen (1966), though he was concerned not with vagueness but with the semantic paradoxes. A glimmering of the idea is found in Russell's suggestion that vagueness is a matter of a one–many relation between words and world (1923, p. 89). Supervaluation theory can be cast in a bivalent metalanguage, and this may have been part of the theory's appeal. However, it can also be cast in a non-bivalent language, as we in effect point out elsewhere (p. 748 and n. 17).

12 Complex predicates must be included, since an appropriate valuation must respect what Fine (1975) has called "penumbral connections" between predicates. Thus a valuation which ensures that "Eve is a female child" is true must also ensure that "Eve is a girl" is true, to respect the intuition that, definitely, nothing satisfies "*x* is a female child and not a girl." Vague names, if there are such things, could be treated just like vague predicates: different valuation functions within a model's family may assign them to different things. But for the moment it will be best to imagine that all assignment functions, or at least all appropriate ones, agree on what they assign to names.

13 If the envisaged definition of validity is retained, this depends upon the assumption that "Def" is not in the object language; see p. 748 and n. 17; and see Williamson (1994).

14 This makes the assumption that arrangement is held constant. It follows that if an *n*-membered collection is a heap, so is any collection with more than *n* members.

15 Thus Wright (1987; 1992) suggests that the notion of there being no sharp boundary can be non-paradoxically expressed using such an operator, whereas removing it leads to immediate paradox.

16 This means that it can do something like the work done by the metalinguistic denial that there is something in the domain of the model which satisfies "$\varphi x \,\&\, \neg\, \varphi x'$."

17 Cf. Williamson (1994, pp. 147–153). For example, although one can validly infer "Def A" from "A," if validity is defined in terms of truth-preservation, this does not guarantee the validity of "If A then Def A": if for some model *M*, some valuation, *f*, "A" is true_f, and for some valuation, *g*, "A" is false_g, then the conditional is false_f and so not true in *M*, so not valid. The supervaluationist might, therefore, prefer to exploit, in his account of validity, a different similarity with the classical definition. His idea is that questions of truth arise only relative to ways of making precise, and he extends this from atomic sentences to complex ones in terms of making the whole complex sentence precise. In this spirit, he might treat validity similarly, defining it in terms of making precise the whole argument. Then, the right thing to say would be that a valid argument is one for which every valuation verifying the premises verifies the conclusion. (As sentence connectives are, in supervaluation theory, valuation-functional but not truth-functional, so with arguments.) It would seem that this would restore the classical character of supervaluational validity. (We are grateful to Dominic Hyde for discussion of this point.)

18

(1)	not True (A)	assumed
(2)	True (A) iff A	(T)
(3)	not A	from (1) and (2) by classical prop. logic
(4)	True (not A) iff not A	(T)
(5)	True (not A)	from (3) and (4) by classical prop. logic
(6)	False (A)	from (5), given that a negation is true iff what it negates is false

This argument will not be persuasive for those theorists who adopt some non-classical logics, e.g., those involving distinct notions of negation.

19 For discussions of the relation between supervaluation theory and higher-order vagueness, see Fine (1975, especially §5), and Williamson (1994, pp. 156–164). On higher-order vagueness more generally, see Heck (1993) and Wright (1992).
20 A classic formulation is Goguen's (1969).
21 "glb" and "lub" abbreviate "greatest lower bound" and "least upper bound" respectively. These are the infinitary analogues of min and max. "$[A^a/v]$" abbreviates "the result of replacing every occurrence of "v" in "Av" by some name not in "Av." Cf. Forbes (1985, p. 174).
22 Supervaluation theory, by contrast, treats the first as neither true nor false and the second as false. This is the problem Fine (1975) calls that of "penumbral connection," and he regards the way in which supervaluation theory handles it as an important merit of the theory.
23 Degrees of truth can be defined within the supervaluational approach: roughly, $[\sigma]$ is the probability of σ being true on a randomly chosen sharpening. Cf. Lewis (1970), Kamp (1975), Edgington (1992), and Williamson (1994, pp. 154–156). However, the two accounts differ on the proper treatment of the logical constants, and they have very different philosophical motivations.
24 See also Horgan (1993), who offers a different non-bivalent account in an article which came to our attention too late to be discussed here. We saw above that supervaluation theory could be modifed so as to allow that its concept of a sharpening is vague (and we mentioned that certain objections would still remain); so it could have featured in both the bivalent and non-bivalent categories of our taxonomy.
25 Early postwar versions of the theory can be found in Cargile (1969) and Campbell (1974). For more recent versions, see Williamson (1992), Sorensen (1988, pp. 217–252), Sperber and Wilson (1986), and Williamson (1994, pp. 185–247).
26 Cf. Hart (1991/1992), though he does not embrace the epistemic view.
27 We thank the editors, Bob Hale and Crispin Wright, for helpful comments on an earlier draft.

References

Barnes, J. 1982. "Medicine, experience and logic." In Barnes *et al.*, 1982, pp. 24–68.
Barnes, J., J. Brunschwig, M. F. Burnyeat, and M. Schofield, eds. 1982. *Science and Speculation*. Cambridge: Cambridge University Press.
Burnyeat, M. F. 1982. "Gods and heaps." In Schofield and Nussbaum, 1982, pp. 315–338.
Campbell, R. 1974. "The sorites paradox." *Philosophical Studies*, 26(3–4): 175–191.
Cargile, J. 1969. "The sorites paradox." *British Journal for the Philosophy of Science*, 20(3): 193–202.
Dummett, M. 1975. "Wang's paradox." Reprinted in *Truth and Other Enigmas*, pp. 248–268. London: Duckworth, 1978.
Edgington, D. 1992. "Validity, uncertainty and vagueness." *Analysis*, 52(4): 193–204.
Fine, K. 1975. "Vagueness, truth and logic." *Synthese*, 30(3–4): 265–300.
Forbes, G. 1985. *The Metaphysics of Modality*. Oxford: Oxford University Press.
Goguen, J. A. 1969. "The logic of inexact concepts." *Synthese*, 19(3–4): 325–373.
Hart, W. D. 1991/1992. "Hat-tricks and heaps." *Philosophical Studies*, 33: 1–24.
Heck, R. G. 1993. "A note on the logic of (higher-order) vagueness." *Analysis*, 53(4): 201–208.
Horgan, T. 1993. "Robust vagueness and the forced-march sorites paradox." In *Philosophical Perspectives*, 8: 159–188.
Kamp, J. A. W. 1975. "Two theories about adjectives." In *Formal Semantics of Natural Language*, edited by E. Keenan, pp. 123–155. Cambridge: Cambridge University Press.
Kamp, J. A. W. 1981. "The paradox of the heap." In *Aspects of Philosophical Logic*, edited by U. Monnich, pp. 225–277. Dordrecht, Netherlands: Reidel.
Lewis, D. 1970. "General semantics." *Synthese*, 22(1): 18–67. Reprinted in his *Philosophical Papers*, vol. 1. Oxford: Oxford University Press, 1983.
Leibniz, G. W. 1961 (1765). *Nouveaux Essais sur l'entendement humain*. Paris: Presses Universitaires de France.

Long, A. A., and D. N. Sedley. 1987. *The Hellenistic Philosophers*, 2 vols. Cambridge: Cambridge University Press.
Mehlberg, H. 1958. *The Reach of Science*. Toronto: Toronto University Press.
Putnam, H. 1983. "Vagueness and alternative logic." *Erkenntnis*, 19(1–3): 297–314.
Putnam, H. 1985. "A quick Read is a wrong Wright." *Analysis*, 45(4): 203.
Read, S., and C. Wright. 1985. "Hairier than Putnam thought." *Analysis*, 45(1): 56–58.
Rolf, B. 1984. "Sorites." *Synthese*, 58(2): 219–250.
Russell, B. 1923. "Vagueness." *Australasian Journal of Philosophy and Psychology*, 1: 84–92. Reprinted in his *Collected Papers*, vol. 9, edited by J. Slater, pp. 145–154. London: Unwin Hyman, 1988.
Schofield, M., and M. C. Nussbaum, eds. 1982. *Language and Logos*. Cambridge: Cambridge University Press.
Sorensen, R. A. 1988. *Blindspots*. Oxford: Clarendon Press.
Sperber, D., and D. Wilson. 1986. *Relevance*. Oxford: Blackwell.
Tye, M. 1994. "Sorites paradoxes and the semantics of vagueness." *Philosophical Perspectives*, 8: 189–206.
van Fraassen, B. 1966. "Singular terms, truth value gaps, and free logic." *Journal of Philosophy*, 63(17): 481–495.
Wiggins, D. R. P. 1980. *Sameness and Substance*. Oxford: Blackwell.
Williamson, T. 1990. *Identity and Discrimination*. Oxford: Blackwell.
Williamson, T. 1992. "Vagueness and ignorance." *Proceedings of the Aristotelian Society*, suppl. vol. 66: 145–162.
Williamson, T. 1994: *Vagueness*. London: Routledge.
Wright, C. 1987. "Further reflections on the sorites paradox." *Philosophical Topics*, 15(1): 227–290.
Wright, C. 1992. "Is higher order vagueness coherent?" *Analysis*, 52(3): 129–139.

Further Reading

Burnyeat, M. F. ed. 1983. *The Skeptical Tradition*. Berkeley: University of California Press.
Frede, M. 1983. "Stoics and skeptics on clear and distinct impressions." In Burnyeat, 1983, pp. 65–93.
Sainsbury, R. M. 1990. *Concepts without boundaries*. Inaugural lecture published by King's College, London.
Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Postscript

AIDAN MCGLYNN

Vagueness and the sorites paradox have continued to be areas of intense focus in the past 20 years. This short appendix cannot possibly do justice to the rich and sophisticated treatments developed in this period, and so instead I offer an overview of some central trends, positions, and points that have emerged.[1]

1 Supervaluationism, Degree Theory, and Epistemicism Revisited

Each of the three principal solutions discussed by Sainsbury and Williamson has continued to be refined and defended. For reasons of space and continuity, I'll focus on the objections raised in Sainsbury and Williamson's discussion.

First, recall Sainsbury and Williamson's objection that adopting supervaluationism forces us to give up certain classically valid rules of inference once we introduce a 'definitely' operator, D, into the object language. Rosanna Keefe has responded that from a supervaluationist perspective, such deviations from classical logic are to be welcomed. D "brings to the object-language the non-classicality of the semantics, since it can be used to capture the fact that some sentence takes the non-classical indeterminate truth-value status" (2000, pp. 178–179), permitting us to express the vagueness of a sentence within the object language. So "the logic of the D operator is appropriately non-classical" (2000, p. 181). Robbie Williams (2008) has questioned whether supervaluationists are committed to logical revisions, arguing that given a plausible assumption about how supervaluationists should understand logical consequence, all of the disputed classical rules are valid even with D in the object language. Moreover, Williams contends that even versions of supervaluationism that are logically revisionary aren't 'damagingly' so, since these logical revisions do not require any revisions to our inferential practices.[2]

Recent work by degree theorists tries to resist the claim that the logics they propose are, as Sainsbury and Williamson allege, 'unintuitive and unmotivated.' For example, if a proposition P receives the value 0.5, then both "$P \wedge \sim P$" and "$P \vee \sim P$" will also be true to degree 0.5, and this has struck many as clearly wrong.[3] John MacFarlane observes that such objections often treat degrees of truth as the *chance* that a proposition is true, and so interpreted it does seem clear that an outright contradiction should receive the value 0. But as MacFarlane notes, "if we are certain that [a contradiction] has degree 0.5, then we will take it to have *no* chance of being completely true" (2010, p. 459). The value 0.5 doesn't reflect maximum *uncertainty* regarding whether a proposition is true, but rather a certain kind of *ambivalence*; it means that one regards it as just as true as false (2010, p. 447).[4] Nicholas Smith (2008, p. 86) also finds the supposedly absurd consequences of degree theories quite intuitive, and he suggests that ordinary speakers frequently agree.[5]

Roy Sorensen has offered a very different version of epistemicism to that defended by Williamson, complaining that the latter makes the unknowability of where the boundary in a sorites series lies too contingent on human epistemic limitations; the kind of ignorance to be explained is '*absolute*' (2001, pp. 13–15, 177). Sorensen's treatment of the sorites paradox is inspired and motivated by his treatment of the 'no-no' paradox. Suppose that you have a sheet of card, and on each side the following statement is printed: 'The statement on the other side of this card is false.' What truth-values should these statements be assigned? We get liar-paradoxical inconsistencies if we try to assign both statements the value true or both false, but the two asymmetrical assignments are consistent. How could one of these statements be true and the other false, though, given their symmetrical situations?

Sorensen's answer is that the no-no is a counter-example to the thesis that every truth has a truthmaker; one of the statements is true and the other false, but *nothing* makes it the case that things are this way around. This explains our absolute ignorance about which statement is the true one; such truths are 'epistemic islands,' since there is no access to them via their truthmakers (2001, p. 175). Sorensen extends this account to explain both how one of the conditionals linking adjacent items in a sorites series could be false while its neighbors are true, and our ignorance of which conditional is the false one.

However, Sorensen's treatment of the no-no paradox has been subject to some serious objections (e.g., López de Sa and Zardini, 2007; 2011; Greenough, 2011). Moreover,

Sorensen's whole account has been subjected to forensic scrutiny by Dorothy Edgington (2005; 2010), and she makes a strong case against it.

2 Quandaries and Intuitionism

A much more radical break with Williamson's epistemicism is Crispin Wright's intuitionism (e.g., 2001; 2003a; 2003b). Wright accepts Williamson's contention that vagueness is at root an epistemic phenomenon, but rather than using this idea to defend classical logic and semantics, Wright takes it as a basis for an argument for intuitionist revisions. Wright treats the paradox as a *reductio ad absurdum* of its major premise, (QP), and so he concludes

$$\sim\forall \mathrm{x}\left(F\mathrm{x} \rightarrow F\mathrm{x}'\right),$$

while refusing to make the (classically valid) step to the 'Unpalatable Existential':

$$\exists \mathrm{x}\left(F\mathrm{x} \wedge \sim F\mathrm{x}'\right).$$

The question is how to motivate such intuitionist restrictions to classical logic.

Borderline cases, according to Wright, present us with *quandaries*, where a proposition *P* presents a quandary to a thinker T just in case: T has considered whether *P*; T does not know whether or not *P*; T does not know any way of coming to know whether or not *P*; T does not know whether there is any way of coming to know whether *P*; and T does not know whether it is even metaphysically possible to know whether *P*. Wright's primary argument for logical revision also requires the thesis that knowledge is closed under known logical entailment (or some refinement of it) and the thesis that propositions making vague predications are known to be subject to a principle of Epistemic Constraint:

$$\left(\mathrm{EC}\right) P \rightarrow \text{it is feasible to know } P$$

Let Bob again be a borderline case of baldness, and let us suppose that the relevant instance of the Law of Excluded Middle,

$$\text{Bob is bald} \vee \text{Bob is not bald}$$

is known. It follows that the following is also known:

$$\text{It is feasible to know that Bob is bald} \vee \text{it is feasible to know that Bob is not bald}$$

But to recognize that borderline cases are quandaries is to recognize that this disjunction is *not* known. Hence, if we know that (EC) holds, we can reject the initial supposition that we know that the Law of Excluded Middle holds in this instance, and this puts pressure on the claim that the Law of Excluded Middle is a logical law (Wright, 2001, p. 66).[6]

I'll briefly mention two objections to Wright's solution. Luca Incurvati and Julien Murzi (2008) contend that a parallel argument can be run against the logical status of the law of non-contradiction, and so Wright's revisionary argument is much more revisionary that he intends, while Sven Rosenkranz (2005) has made a plausible case that Wright's quandaries conception of borderline cases is in fact incoherent (see also Rosenkranz, 2003, and Wright, 2003b).

3 Dialetheism as a Unified Solution

Graham Priest has revived Bertrand Russell's suggestion that the paradoxes of self-reference – the Liar paradox, Russell's paradox, the Burali-Forti paradox, and so on – share a common underlying structure, and so should receive a unified solution (e.g., Priest, 2002). Priest classifies these paradoxes as *inclosure* paradoxes, since they fit *the inclosure schema*. We have an inclosure paradox just when there are monadic predicates φ and Ω and a one-place function δ such that we can construct apparently sound arguments for the following claims:

1. There is a set Ω such that $\Omega = \{x : \varphi(x)\}$ (Existence)
2. If $X \subseteq \Omega$,
   a. $\delta(X) \notin X$ (Transcendence)
   b. $\delta(X) \in \Omega$ (Closure)[7]

Consider the Burali-Forti paradox, which shows that there cannot be a set of all ordinal numbers on pain of contradiction. Here's how to represent the paradoxical reasoning as an inclosure paradox (Priest, 2010, p. 70). $\varphi(x)$ is 'x is an ordinal,' and so Ω is the set of ordinals, which we'll take to be defined as von Neumann ordinals (so that each ordinal is taken to be the well-ordered set of all smaller ordinals). Let $\delta(X)$ be the least ordinal greater than every member of X. $\delta(X)$ satisfies Transcendence and Closure by definition. In the special case where $X = \Omega$, we get the contradiction $\delta(\Omega) \in \Omega$ and $\delta(\Omega) \notin \Omega$: intuitively, the least ordinal greater than every member of the set of ordinals both is and is not a member of the set of ordinals.

To be faced with an inclosure paradox, it's not required that the arguments for Existence, Transcendence, and Closure ultimately prove sound or even valid. However, *dialetheists* like Priest argue that the best response to the paradoxes of self-reference involves taking the apparent soundness at face value, and accepting the resulting contradictions. Triviality is avoided by adopting a *paraconsistent* logic: one lacking the principle of explosion (*ex falso quodlibet*).

Priest (2010) has recently argued that the sorites paradox is also an inclosure paradox, and so since the paradoxes of self-references are to be given a dialetheic solution, so should the sorites.[8] Let F be our vague predicate, and we'll assume it obeys (QP):

$$\text{(QP)}\ \forall x\ \left(Fx \supset Fx'\right)$$

Let $\varphi(x) = Fx$, so that $\Omega = \{x : Fx\}$. This gives us existence. Ω is a proper subset of the series, since the last item in the series is not F. If $X \subseteq \Omega$, then X must also be a proper subset of the series. So there must be a first item in the series not in X. Let this be $\delta(X)$. By definition $\delta(X) \notin X$, giving us Transcendence. But also by definition, $\delta(X)$ is the successor of an item in $X \subseteq \Omega$. So by (QP), $F\delta(X)$, in which case $\delta(X) \in \Omega$. And so we have Closure too. When we let $X = \Omega$, the resulting contradiction is that $\delta(\Omega) \in \Omega$ and $\delta(\Omega) \notin \Omega$, or as Priest puts it (2010, p. 71), the first item in the series that is not F is F. Intuitively, the clash here is between (QP) and the thought that, since the set of Fs is a proper subset of the entire series, then given the ordering of the series there must be a first item in the series *not* in that set.

Note that while Priest thinks that this soritical inclosure reasoning is sound (and so requires dialetheism), he doesn't think that "sorites arguments themselves" are sound (2010, pp. 72, 79–80); that's to say, he doesn't accept the argument from the premise that the first

item in a sorites series is *F* and (QP) to the conclusion that every item in the series is *F*. Different dialetheists have addressed this point in different ways; according to Priest, sorites arguments are blocked by the fact that in a paraconsistent setting, *modus ponens* isn't valid for any conditional that renders (QP) defensible.

Many philosophers find dialetheism deeply unattractive, and importantly, there are specific concerns that can be raised to its application to the sorites. The most pressing begins with consideration of Curry's paradox; one can derive an arbitrary proposition *P* as a theorem, starting with a self-referential conditional that says

If this statement is true then *P*,

and appealing only to the standard introduction and elimination rules for the conditional and Tarksi's disquotational schema (or something like it).[9] Unless blocked, Curry's paradox renders our logic trivial, and crucially it does so without requiring us to pass through a contradiction, and so without an appeal to the principle of explosion. J. C. Beall (2014a; 2014b) has argued at length that Curry's paradox has at least as much of a claim to count as an inclosure paradox as the sorites, but it can't receive a dialetheic solution, putting pressure on the argument for a dialetheic treatment of the sorites.[10]

4 Contextualism and Interest-Relativity

According to *contextualism* about vagueness, vague expressions are characteristically context sensitive in ways that go beyond what we standardly recognize (such as the fact that whether someone counts as 'tall' in a conversational context depends on the relevant comparison class), and these hidden dimensions of context sensitivity explain the allure of (QP), even though it is not true. Though the details vary, the general idea is roughly that when any two sufficiently similar items in a given sorites series are considered, the interpretation of the vague predicate shifts (if it needs to) so that the first falls under it only if the second does too. So whenever we consider whether a pair of adjacent items in a sorites series is a counter-example to the major premise of the relevant paradox, that very act ensures that it is not (e.g., Raffman, 1994). Crucially, though, there is no single context in which every instance of the major premise is true. Contextualism might offer a way to defend classical logic and semantics; however, it is not wedded to a classical picture (e.g., Soames, 1999, ch. 7; 2002).[11]

Jason Stanley takes contextualism to be committed to the claim that vague expressions are indexicals (2003, p. 271), and he contends that this renders the contextualist's explanation of why (QP) is attractive insufficiently general. First, he argues that indexicals do not shift their interpretation under verb phrase ellipsis. Consider an example:

(1) John likes me, and Bill does too.

As Stanley notes, "[t]here is no available interpretation of (1) in which John and Bill are said to like different people" (2003, p. 271). Stanley then constructs a sorites series that exploits verb phrase ellipsis. He imagines a series of arrangements of sand that starts with a clear heap, and where each successive arrangement contains one less grain. Pointing at each arrangement of sand in turn:

> If that$_1$ is a heap, then that$_2$ is too, and if that$_2$ is, then that$_3$ is, and if that$_3$ is, then that$_4$ is, … and then that$_n$ is. (2003, p. 272)

The explanation of why we find each of these conjuncts compelling cannot be that shifts in the interpretation of 'heap' ensure that each conjunct is true as we consider it, Stanley concludes, since there are no such shifts.

Jonathan Ellis (2004) has disputed Stanley's claim about indexicals, while Diana Raffman (2005) contends that contextualists should deny that vague expressions are indexicals. Delia Graff Fara (2000) offers an account that's closely related to contextualism, and which allows her to offer a similar treatment of the sorites paradox while evading Stanley's objection. According to Fara, vague expressions are no more context sensitive than ordinary thought, but "the semantics of vague expressions renders the truth-conditions of utterances containing them sensitive to our interests" (2000, p. 49). Here the role our interests play is not in fixing the interpretation of an expression like 'heap' in a context; rather, they play a role in determining whether the (relatively) context-invariant standards for being a heap are met.[12]

However, Stanley's objection is a version of a more general worry, namely that (QP) remains compelling even when we hold fixed all of the features relative to which any boundaries in the associated sorites series supposedly shift. If that's right, then the general explanation of why we are taken in by (QP) cannot be that hidden shifts ensure that we never confront a counter-example to it.[13]

Notes

1 To keep things manageable, I bracket issues concerning higher-order vagueness, though I offer references in the notes and in the suggested further reading.

2 See McGee and McLaughlin (1994) and Asher, Dever, and Pappas (2009) for other developments of broadly supervaluationist approaches that claim to preserve classical logic, and see Keefe (2000, ch. 8) and Asher, Dever and Pappas (2009) for supervaluationist treatments of higher-order vagueness.

3 I leave aside versions of degree theory on which the connectives are not truth-functional and so which lack this result (e.g., Edgington, 1997); see Smith (2008, pp. 263–264) for discussion.

4 Unlike other defenders of a degree-theoretic approach to vagueness, MacFarlane doesn't think that this approach solves the sorites paradox.

5 A recent empirical study by David Ripley (2011) partially bears Smith's claim out. For degree-theoretic treatments of higher-order vagueness see, e.g., Smith (2008, pp. 304–315) and MacFarlane (2010).

6 This sketch of Wright's arguments ignores some subtleties addressed in his 2001. Wright also explores some arguments that don't appeal to (EC). See Wright (2003a, pp. 102–104; 2003b, pp. 472–473).

7 Actually, this is Russell's schema (Priest, 2002, p. 129), which Priest's inclosure schema generalizes. The difference won't matter here.

8 See also Weber (2010), and see Weber *et al.* (2014, p. 813, fn. 1) for some background to Priest's discussion.

9 Curry (1942); see Cook (2013, pp. 71–77) for a more accessible presentation.

10 See Weber *et al.* (2014) for critical discussion of Beall. An alternative unified solution to the paradoxes (potentially including Curry's paradox) involves placing restrictions on the *structural* rules built into classical logic (as opposed to the *operational* rules that govern the various logical constants). For such approaches to the sorites paradox, see, e.g., Ripley (2013) and Zardini (2008).

11 See also Shapiro (2006) (though his treatment of the sorites paradox is rather different to that sketched in the text), and see Keefe (2003) and Greenough (2005) for criticism.

12 For critical discussion of Fara see Stanley (2003, pp. 277–279); but see Fara (2008) for a reply. See also Heck (2003, pp. 119–120).

13 This general worry is critically discussed in Åkerman and Greenough (2010a). Keefe (2007) offers a battery of further arguments against contextualism: see Åkerman and Greenough (2010b) for discussion.

References

Åkerman, J., and P. Greenough. 2010a. "Hold the context fixed – vagueness still remains." In Dietz and Moruzzi, 2010, pp. 275–288.

Åkerman, J., and P. Greenough. 2010b. "Vagueness and non-indexical contextualism." In *New Waves in Philosophy*, edited by S. Sawyer, pp. 8–23. Basingstoke: Palgrave Macmillan.

Asher, N., J. Dever, and C. Pappas. 2009. "Supervaluations debugged." *Mind*, 118(472): 901–933.

Beall, J. C., ed. 2003. *Liars and Heaps*. Oxford: Oxford University Press.

Beall, J. C. 2014a. "Finding tolerance without gluts." *Mind*, 123(491): 791–811.

Beall, J. C. 2014b. "End of inclosure." *Mind*, 123(491): 829–849.

Cook, R. 2013. *Paradoxes*. Cambridge and Malden: Polity Press.

Curry, H. 1942. "The inconsistency of certain formal logics." *Journal of Symbolic Logic*, 7(3): 115–117.

Dietz, R., and S. Moruzzi, eds. 2010. *Cuts and Clouds: Vagueness, its Nature, and its Logic*. Oxford: Oxford University Press

Edgington, D. 1997. "Vagueness by degrees." In *Vagueness: A Reader*, edited by R. Keefe and P. Smith, pp. 295–316. Cambridge, MA: MIT Press.

Edgington, D. 2005. "The mystery of the missing boundary: *Vagueness and Contradiction* by Roy Sorensen." *Philosophy and Phenomenological Research*, 71(3): 704–711.

Edgington, D. 2010. "Sorensen on vagueness and contradiction." In Dietz and Moruzzi, 2010, pp. 91–106.

Ellis, J. 2004. "Context, indexicals and the sorites." *Analysis*, 64(4): 362–364.

Fara, D. G. 2000. "Shifting sands: an interest-relative theory of vagueness." *Philosophical Topics*, 28(1): 45–81. Originally published under the name Delia Graff.

Fara, D. G. 2008. "Profiling interest relativity." *Analysis*, 68(4): 326–335.

Greenough, P. 2005. "Contextualism about vagueness and higher-order vagueness." *Proceedings of the Aristotelian Society*, suppl. vol. 79: 167–190.

Greenough, P. 2011. "Truthmaker gaps and the no-no paradox." *Philosophy and Phenomenological Research*, 82(3): 547–563.

Heck, R. 2003. "Semantic accounts of vagueness." In Beall, 2003, pp. 106–127.

Incurvati, L., and J. Murzi. 2008. "How basic is the basic revisionary argument?" *Analysis*, 68(4): 303–309.

Keefe, R. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Keefe, R. 2003. "Context, vagueness, and the sorites: comments on Shapiro." In Beall, 2003, pp. 73–83.

Keefe, R. 2007. "Vagueness without context change." *Mind*, 116(462): 275–292.

Keefe, R., and P. Smith, eds. 1997. *Vagueness: A Reader*. Cambridge, MA: MIT Press.

López de Sa, D., and E. Zardini. 2007. "Truthmakers, knowledge, and paradox." *Analysis*, 67(3): 242–250.

López de Sa, D., and E. Zardini. 2011. "No-no. Paradox and consistency." *Analysis*, 71(3): 472–478.

MacFarlane, J. 2010. "Fuzzy epistemicism." In Dietz and Moruzzi, 2010, pp. 438–464.

McGee, V., and B. McLaughlin. 1994. "Distinctions without a difference." *Southern Journal of Philosophy*, 33(suppl.): 203–251.

Priest, G. 2002. *Beyond the Limits of Thought*, 2nd edn. Cambridge: Cambridge University Press.

Priest, G. 2010. "Inclosures, vagueness, and self-reference." *Notre Dame Journal of Formal Logic*, 51(1): 69–84.
Raffman, D. 1994. "Vagueness without paradox." *Philosophical Review*, 103(1): 41–74.
Raffman, D. 2005. "How to understand contextualism about vagueness: reply to Stanley." *Analysis*, 65(3): 244–248.
Raffman, D. 2014. *Unruly Words: A Study of Vague Language*. Oxford: Oxford University Press.
Ripley, D. 2011. "Contradictions at the borders." In *Vagueness in Communication*, edited by R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz, pp. 169–188. Heidelberg: Springer.
Ripley, D. 2013. "Revising up: strengthening classical logic in the face of paradox." *Philosophers' Imprint*, 13(5): 1–13.
Rosenkranz, S. 2003. "Wright on vagueness and agnosticism." *Mind*, 112(447): 449–463.
Rosenkranz, S. 2005. "Knowledge in borderline cases." *Analysis*, 65(285): 49–55.
Shapiro, S. 2006. *Vagueness in Context*. Oxford: Oxford University Press.
Smith, N. 2008. *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.
Soames, S. 1999. *Understanding Truth*. Oxford: Oxford University Press.
Soames, S. 2002. "Replies." *Philosophy and Phenomenological Research*, 65(2): 429–452.
Sorensen, R. 2001. *Vagueness and Contradiction*. Oxford: Oxford University Press.
Stanley, J. 2003. "Context, interest relativity and the sorites." *Analysis*, 63(280): 269–280.
Weber, Z. 2010. "A paraconsistent model of vagueness." *Mind* 119(476): 1025–1045.
Weber, Z., D. Ripley, G. Priest *et al.* 2014. "Tolerating gluts." *Mind*, 123(491): 813–828.
Williams, R. 2008. "Supervaluationism and logical revisionism." *Journal of Philosophy*, 105(4): 192–212.
Wright, C. 2001. "On being in a quandary: relativism, vagueness, logical revisionism." *Mind*, 110(437): 45–98.
Wright, C. 2003a. "Vagueness: a fifth column approach." In Beall, 2003, pp. 84–105.
Wright, C. 2003b. "Rosenkranz on quandary, vagueness, and intuitionism." *Mind*, 112(447): 465–474.
Zardini, E. 2008. "A model of tolerance." *Studia Logica*, 90(3): 337–368.

Further Reading

Cook (2013) is an accessible introduction to paradoxes in general, and it includes a chapter on the sorites as well as coverage of many other topics discussed here. Keefe and Smith (1997) is an anthology of now classic papers on vagueness, while Beall (2003) and Dietz and Moruzzi (2010) are collections of recent work on the topic that both include important discussions of all of the issues covered here, as well as several key papers on higher-order vagueness. Important recent monographs on the sorites include Keefe (2000), Sorensen (2001), Shapiro (2006), Smith (2008), and Raffman (2014).

29

Time and Tense

BERIT BROGAARD

1 Introduction

Two of the main debates in philosophy of language concerning time and tense are the debate about the semantics of the tenses in the English language and the debate over whether propositions can be transiently true or false as opposed to always being eternally true or false. The latter quarrel is also known as the 'temporalism–eternalism debate.' Given standard semantics, the two debates are not logically independent, as we will see. Those who believe propositions are eternally true or false needn't treat the tenses as operators. Their opponents, on the other hand, appear to be committed to an operator theory of the tenses, given a standard semantic framework. In this chapter I will focus primarily on these two debates.

There are many other debates about time in philosophy of language. For example, there is a question about how we can best account for the cognitive significance of claims like 'Today is February 5, 2014' or *de se* beliefs such as John's belief that it's 3 o'clock now. On a standard account of propositional content, 'Today is February 5, 2014' expresses the proposition that February 5, 2014 is February 5, 2014. But the latter is trivially true and not something anyone needs to discover empirically. Finding out that today is February 5, 2014, on the other hand, can be an important empirical discovery. If you got married on February 5, knowing that today is February 5, 2014, may help you not getting into trouble with your spouse. John's *de se* belief that it's 3 o'clock now presents its own problems, because this belief clearly isn't the same as the belief that it's 3 o'clock at 3 o'clock. So, beliefs cannot simply be relations to the propositions of standard semantics. Although these other debates about time are interesting, they are not exclusively about time. Accounting for the cognitive significance of claims like 'I am Brit' or *de se* beliefs such as John's belief that he is the shopper who is making a mess, presents the same challenges to standard semantics as the analogous examples involving time. For this reason I shall not deal with these issues except in passing.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

At the end of the chapter I will briefly look at the relevance of debates about tense and eternalism/temporalism to metaphysical debates about time. I will argue that the debates in philosophy of language are not logically independent of the debates in metaphysics.

2 Temporalism versus Eternalism

2.1 Times in Propositions versus Time Neutrality

Mark Richard (1981) calls the thesis that propositions are unable to change their truth-values over time 'eternalism' and the opposing view 'temporalism.' Temporalism is committed to the view that either some propositional attitudes have temporal propositions as their objects or sentences that lack time adverbials (e.g., 'now,' 'when John was born,' 'at 2 p.m. July 6, 2005') express, relative to a context of use, temporal propositions. Propositions of this sort may vary in truth-value over time. For example, the proposition expressed, relative to a context of use, by 'This tree is covered with green leaves' may be true in the summer but false in the winter.

Whether eternalism or temporalism is correct will depend on which of the two views (if any) best accounts for the features that propositions have traditionally been said to have. Traditionally, propositions have been thought to play a number of distinct theoretical roles: Propositions are (i) the semantic values of truth-evaluable sentences, (ii) the objects of the attitudes (e.g., belief, doubt, hope, wish, and so on), (iii) the objects of agreement and disagreement, (iv) what is transferred or shared when people communicate successfully, and (v) the contents intensional operators operate on (e.g., modal operators or tense operators). It may turn out that neither temporalism nor eternalism can ensure that propositions play all of these roles. The verdict (if any) will then depend on which view allows us to preserve most of these roles, or most of the roles that we regard as most important.

One feature that has traditionally been regarded as important is that propositions are the entities that modal operators and tense operators operate on. Taking something like this claim for granted, David Kaplan (1989) has offered a now well-known argument for the view that there are temporal propositions, that is, propositions that are capable of changing their truth-values over time. The argument runs as follows.

Kaplan's Argument
(A) There are non-redundant tense operators in English.
(B) Tense operators operate on propositions.
(C) Tense operators that operate on eternal propositions are semantically redundant.
(D) Hence, tense operators operate on temporal propositions.
(E) Hence, there are temporal propositions.

Kaplan takes premise (A) to be empirically evident. In Kaplan's opinion, premise (B) is relatively innocent as well. It should be said, however, that Kaplan does not insist on the term 'proposition.' In fact, his use of scare quotes reflects his 'feeling that this is not the traditional notion of a proposition' (1989, p. 503). However, the claim that tense operators operate on content rather than, say, linguistic meaning is an important corollary of the theory of Kaplan's "Demonstratives." Premise (C) is the key premise of the argument, as far as Kaplan is concerned. The argument for premise (C) runs as follows. Consider a sentence containing a past tense operator, such as:

(1) It has been that John is a firefighter.

The past tense operator 'it has been that' shifts the time feature of the circumstance of evaluation at which the content of sentence (1) is evaluated from the time of speech to some time in the past. If, however, the content of 'John is a firefighter' were eternal, it would have the same truth-value with respect to any time of evaluation. So 'It has been that John is a firefighter' would have the same truth-value as the operand sentence 'John is a firefighter,' which is to say that 'it has been that' would be semantically redundant. The argument is nicely summarized in this footnote from "Demonstratives":

> Technically, we must note that intensional operators must, if they are not to be vacuous, operate on contents which are neutral with respect to the feature of circumstance the operator is interested in. Thus, for example, if we take the content of S to be [eternal], the application of a temporal operator to such a content would have no effect; the operator would be vacuous. Furthermore, if we do not wish the iteration of such operators to be vacuous, the content of the compound sentence containing the operator must again be neutral with respect to the relevant feature of circumstance. This is not to say that no such operator can have the effect of *fixing* the relevant feature and thus, in effect, rendering subsequent operations vacuous; indexical operators do just this. It is just that this must not be the general situation. A content must be the *kind* of entity that is subject to modification in the feature relevant to the operator. (Kaplan, 1989, pp. 503–504, n. 28)

Intensional operators must operate on contents whose truth-value varies with the feature shifted by the operator. Otherwise, they are semantically redundant. Since the truth-values of eternal propositions do not vary with time, tense operators that operate on eternal propositions are semantically redundant. Hence, if there are non-redundant tense operators in the language, then they operate on temporal propositions.

Premises A and B, however, are not as innocent as they may at first seem. As we will see below, premise A, *viz*. the claim that there are tense operators in the English language, has been disputed by both philosophers and linguists. It is a minority view among linguists today but is still fairly commonly accepted among philosophers. One alternative to the operator view is the quantifier view, defended by, for example, Jeff King (2003). On this view, 'it was the case that it was raining in St Louis' expresses the proposition that there is a time t that is earlier than the time of utterance t^*, and it is raining in St Louis at t. We will return to the debate about the tenses below.

Premise B, *viz*. the claim that tense operators operate on propositions, has also been rejected by several philosophers, including Michael Dummett (1991), Nathan Salmon (1989), and Jason Stanley (1997a; 1997b). Although these thinkers accept that there are tense operators in the English language, they deny that they operate on propositions. They distinguish between the assertoric content and the compositional semantic value of a sentence. Tense operators, they say, operate on the compositional semantic value, whereas assertoric contents serve as the objects of belief and the contents of utterances. Assertoric contents, they say, serve as propositions, whereas compositional semantic values do not serve as propositions. The former are eternally true or false, whereas the latter can take on different truth-values at different times. So, drawing this distinction allows its defenders to advocate for eternalism without having to reject the claim that there are tense operators in the English language. Although the distinction between assertoric content and compositional semantic values is not uncontroversial (see Brogaard, 2012, ch. 6), a knock-down argument against this view has still to be provided. So, Kaplan's argument cannot currently be considered a successful argument for temporalism.

In his dissertation, Clas Weber (2013, ch. 2) has provided a variation on Kaplan's argument that he says avoids the problematic intensional assumptions of Kaplan's argument. He calls this argument 'the substitution argument.' It starts with the observation that for eternalists sentences without an explicit time specification express the same eternal propositions as the sentences in which the time specification is made explicit. So, 'it is raining in Canberra' and 'it is raining in Canberra at t^*,' where t^* is the time of utterance, are semantically equivalent for the eternalist. He calls pairs of sentences of this kind 'eternalization pairs.' Now consider the following eternalization pair:

(2) It is raining in Canberra.
(3) It is raining in Canberra on the 22nd of August 2010 at 2:36 p.m.

Because (2) and (3) are an eternalization pair, they are semantically equivalent as far as the eternalist is concerned. Despite this, (2) and (3) cannot be interchanged *salva veritate* within the temporal construction *It is always the case that*.

(4) It is always the case that it is raining in Canberra.
(5) It is always the case that it is raining in Canberra on the 22nd of August 2010 at 2:36 p.m.

Weber argues that (4) is false, whereas (5) is not. (4) implies that Canberra, like Seattle, always requires you to bring an umbrella when you go out. (5), on the other hand, implies that it is eternally true that it is raining on a particular day in Canberra. Because substitution fails in the temporal context, Weber says, (2) and (3) cannot be semantically equivalent after all. So, eternalism is false.

However, I don't think Weber is right that the substitution argument does not rest on the intensional assumptions of Kaplan's argument. Kaplan's argument relies on, for example, the assumption that there are tense operators in the language (e.g., 'It is always the case that'). So does the substitution argument. It is clear that the argument does make this assumption, otherwise substitution would be illicit for different reasons. For example, substitution of 'it is raining at t_1' for 'it is raining' is illicit in the following context if the tenses are quantifiers rather than operators:

(6) It was the case that it is raining.

If this is analyzed as 'there is a time t such that t is earlier than the time of speech, and it is raining at t,' then substituting 'it is raining at t_1' for 'it is raining' is clearly illicit. The same holds for 'it is always raining.' If this is analyzed as 'for all times t, it is raining at t,' then it is illicit to substitute 'it is raining at t_1' for 'it is raining.' So, the substitution argument does, in fact, rest on the assumption that there are tense operators in the English language. Only these kinds of operators would make the substitution illicit for reasons relevant to the temporalism–eternalism debate.

But if 'it was the case that' is a tense operator, then it is open to defenders of the distinction between assertoric content and compositional semantic values to argue that the reason we cannot embed (2) and (3) within the scope of this temporal operator is that temporal operators operate on compositional semantic values. So, when (2) is embedded under 'it is

always the case that,' then it has a different content than when it is not embedded. This explains why substitution is elicit. The substitution argument thus does not seem to fare better than Kaplan's original argument.

2.2 Richard's Argument

Another well-known argument central to the temporalism–eternalism debate was set forth by Mark Richard (1981). The argument is supposed to show that there are obviously invalid arguments that would come out valid if temporalism were true. So temporalism is false. Here is one such apparently invalid argument:

(A)
Mary believed that Nixon was president.
Mary still believes everything she once believed.
Therefore, Mary believes that Nixon *is* president.

According to Richard, "this argument is not a valid argument in English. As speakers of English use sentences such as [premise 1] and [premise 2], [the conclusion] simply does not follow from them" (1981, p. 4). Or, as Salmon puts it, "such an inference is an insult not only to Mary but also to the logic of English, as it is ordinarily spoken" (1989, p. 345). Yet, says Richard, the temporalist must regard (A) as valid. On behalf of the temporalist, Richard assigns the following metalinguistic truth-conditions to (A):

$$\exists p \exists t(t < t^* \,\&\, p = [P_n] \,\&\, Bmpt)$$
$$\forall p(\exists t(t < t^* \,\&\, Bmpt) \rightarrow Bmpt^*)$$
$$\exists p(p = [P_n] \,\&\, Bmpt^*)$$

(p ranges over propositions, '$<$' means 'is earlier than,' t^* is the time of speech, m is a constant that refers to Mary, and $[P_n]$ is the temporal proposition that Nixon is president). The first premise is true iff there is a time t such that t is earlier than the time of speech t^*, and a proposition p such that p is *Nixon is president* and at t Mary believes that p. The second premise is true iff for all propositions p, if there is a time t that is earlier than the time of speech t^* and Mary believes that p, then at the time of speech t^* Mary believes that p. The conclusion is true iff there is a proposition p such that p is *Nixon is president*, and at the time of speech t^* Mary believes that p. But this is valid. So the temporalist is committed to the validity of an apparently invalid argument.

The eternalist is not so committed. For the eternalist takes the first premise to mean that there is a time t such that t is earlier than the time of speech t^*, and Mary believes at t that Nixon is president *at t*. From this and the assumption that Mary still believes everything she once believed it does not follow that Mary believes at t^* that Nixon is present at t^*. In other words, the difference between the verdicts of temporalism and eternalism is that temporalism takes the objects of beliefs to be temporally neutral, whereas eternalism takes them to be temporally specified.

As I have argued on previous occasions, the main problem with this argument is that there are structurally analogous arguments that seem to us to be valid, which is what temporalism would predict (see, e.g., Brogaard, 2012, ch. 2). Here are a few examples:

(B)
John will be thinking that Mary is hungry.
Everything John will be thinking he is thinking now.
Therefore, John is thinking that Mary is hungry.

(C)
Yesterday John believed that Arnold Schwarzenegger was the president of the United States.
Today John believes whatever he believed yesterday.
Therefore, John believes that Arnold Schwarzenegger is the president of the United States.

(D)
Yesterday John pretended that he was a famous actor.
Now he is pretending that same thing again.
Therefore, John is pretending that he is a famous actor.

(E)
Yesterday John dreamed he was the president of the United States.
Now he is dreaming the same thing as yesterday.
Therefore, John is dreaming that he is the president of the United States.

Unlike eternalism, temporalism correctly predicts that these arguments are valid (the verdict of untutored informants). Although these arguments are structurally analogous to Richard's argument (A), our intuitions differ in these cases. So, our intuitions regarding argument (A) cannot be used to establish that temporalism is false. The temporalist might say that what goes wrong in argument (A) is that the conclusion is so outrageous ('an insult to Mary') that we automatically reject it, despite the fact that it follows from the premises. If the conclusion were less outrageous, perhaps our intuitions would be less strong. We might test that hypothesis by considering the following argument:

(F)
Mary believed that Obama was president.
Mary still believes everything she once believed.
Therefore, Mary believes that Obama *is* president.

Most people asked about the status of this argument seem to think it is perfectly fine, or at least it does not seem obviously invalid to them. So, it may be that argument (A) seems outrageous because the conclusion is outrageous. It is hard to envisage that Mary could be that stupid.

2.3 Belief Retention

Richard (1981) also argues that the temporalist has a problem accounting for belief retention. Consider the following example of belief retention:

(G)
I, Mary, believed that Nixon was up to no good in the White House, and I still believe that.
Therefore, I, Mary, believe that Nixon is up to no good in the White House.

Intuitively, (G) is invalid. Yet, says Richard, the temporalist is committed to its validity. For, given temporalism, the premise is true iff there is a time *t* such that *t* is earlier than the time of speech t^* and Mary believes at *t* that Nixon is up to no good in the White House, and at t^* Mary still believes that Nixon is up to no good in the White House. From this, of course, it follows that at t^* Mary believes that Nixon is up to no good in the White House.

Eternalism, on the other hand, is not committed to this result. According to the eternalist, the objects of the attitudes are eternal. So, the premise is true if and only if there is a time *t* such that *t* is earlier than the time of speech t^* and Mary believes at *t* that Nixon is up to no good in the White House at *t*, and at t^* Mary still believes that Nixon is up to no good in the White House at *t*.

Richard (1981, p. 6) considers the possibility of the temporalist offering an alternative account of belief retention. On this view, "to retain a belief is *not* to continue to believe the very same proposition. Rather, it is to believe a proposition related in some special way to the proposition originally believed" (1981, p. 6). To believe what one once believed is to believe that it was the case that what one once believed obtains. For example, if Mary once believed that Nixon is president, and she retains this belief, then she now believes that Nixon was president. This move would block Richard's argument. For from the assumption that Mary once believed that Nixon is president but now believes that Nixon *was* president, it does not follow that she believes that Nixon *is* president. Richard thinks this account of belief retention is unacceptable.

However, as it turns out, (G) is not a good way of refuting temporalism for the same reason that (A) is not a good way to refute temporalism. It may simply be that we find Mary's claim that she believes that Nixon is still in the White House so stupid that we implicitly reject the entire reasoning process, in spite of it being valid.

What Richard does not realize is that it is a much greater challenge to come up with an adequate account of how belief is retained over time if the objects of belief are eternal propositions. The problem for the eternalist is that we rarely retain belief for the long term by remembering the same eternal proposition. Presumably, when we store a belief about a present occurrence, we store it as a past-tensed proposition. For example, if at 15:13 on January 5, 2010, I see a red car leave a crime scene, I will likely store the information as a past-tensed proposition; for example, I may store the information in the form *it was the case on that day where I observed the horrible crime that a red car left the crime scene*.

Because eternalists are committed to the claim that all propositions make reference to a time, they cannot account for this way of storing information. They might say that the information I store has the form 'there is a time *t* such that *t* is prior to or identical to t^*, and Brit observes a terrible crime just before *t* and a red car is leaving the crime scene at *t*,' where 't^*' refers to the time at which the belief information is stored, for example, 15:13 on January 5, 2010. But surely this is not the kind of information that is likely to get stored. To store this kind of information the brain would need to be in a position to track the time precisely at the time of storage. It is just plainly implausible that the brain would have tracking powers like that. If, on the other hand, the brain stores a temporal proposition, then belief retention consists in continuing to stand in the belief relation to the same temporal proposition. So, eternalism is false.

The eternalist may insist that they have a way of dealing with this sort of case. What I store in my hippocampus is not a proposition that refers to a specific time but rather a proposition that quantifies over times. I observe the crime and see the red car escape and then I form the belief that there is a time *t* such that Brit observes a terrible crime just before

t and sees a red car escape at *t*. While this gets around the problem of how the brain stores information about specific times on the basis of observations of a scene with no clocks, it runs into trouble of a different kind. When I retrieve the stored information, my retrieved memory can be true even if I never observed a crime in my life. It could be true if I were to observe a red car escape a crime scene 10 years from now. Our ordinary life experiences tell us that it is unlikely that I falsely remember the details about an event that then occurs in the same way 10 years later. But memories need not be very detailed. If I am told at time *t* that I got an A for my essay about Columbus, this may be all I am able to recall later about the situation in which I learned this fact and about the essay. But if my brain stores the information that there is a time at which I get an A for my essay about Columbus, then what I recall could be true, even if the only essay I ever wrote about American history was about Lewis and Clark. It would be true if I were to go back to school later and were to earn an A for my essay about Columbus. Belief information clearly is not stored in memory in this kind of tense-neutral way. Information about the past is stored for the long term in a past-oriented way.

We don't always continue to believe a proposition by storing the information in storage memory. Sometimes I continue to believe something over time without storing the information in storage memory at all. This is the case when we keep information available in working memory. For example, if I want to call you, I may look up your phone number in the phonebook. As phone books are reliable sources of information, I rationally come to believe that your phone number is, say, 283-1759. I can keep this information available in working memory for the few minutes it takes me to find my phone and dial the number. The information I keep available in working memory for the few minutes it takes me to find my phone and dial the number is hardly indexed to a specific time. I don't continue to believe that your number is 283-1759 by believing that your number is 283-1759 at 15:00 on July 5, 2010, that your number is 283-1759 at 15:01 on July 5, 2010, and so on. The information I keep in mind is just the non-indexed information that your phone number is 283-1759. For as long as I keep that information available in my mind, I stand in a belief relation to the information. So, it is possible to stand in a belief relation to temporal content. This is in conflict with eternalism.

In general, it seems that information can be retained over time in two different ways. One can retain it in the past tense, or in the present tense. Information about occurrences typically is stored in the past tense, whereas information about things that continue to exist over time may be stored in the present tense. When I saw the red car leave the crime scene, that's an occurrence, and the information is therefore stored in the past tense together with some temporal markers. Since phone numbers exist over time and do not change very quickly, information about phone numbers may be stored as a present-tensed proposition that does not make reference to a time.

The duality in how we retain belief is reflected in the language we use to talk about it. If I say "Four years ago I believed that John was a firefighter, and I still believe it," then I can either mean that I still believe that John is a firefighter or that I still believe that he was a firefighter then. But the standard version of eternalism cannot account for the duality in the meaning of these sentences. The standard version is required to interpret the second clause as being a time-indexed claim about John four years ago. At best this captures the second reading. The other reading is unaccounted for.

The following examples shed further light on the difficulty that the duality in the meaning of these sentences presents for eternalism (Brogaard, 2012, ch. 2):

DECEIT

WIFE: When I married John I thought he was a police officer. Thirty years later I still believed he was a police officer. Turns out that he was fired two years into our marriage.

LOST LOVE

FRIEND: Yes, Barbara did love you 10 years ago. So you were right back then. But you still believe that she loves you, don't you Peter?

DEFENSE

STUDENT: I think my dissertation is done.

SUPERVISOR: You do? Well, I think you are wrong. Work on it for a few more weeks. Then read it again. If you *still* think that it's done, then we'll talk.

It is important to note here that these cases are about still believing that something is the case rather than believing that something still is the case. So, in LOST LOVE, for example, the friend claims that Peter still believes that Barbara loves him. The latter claim is distinct from the claim that Peter believes that Barbara still loves him. The two claims may be closely related but it is the former construction I am interested in here.

Eternalism holds that what we believe when we believe something that is not in the past tense is a time-indexed proposition. So, in the envisaged example outlined in DECEIT, the wife's original belief has the propositional content *my husband is a police officer at t*, where *t* is some time 30 years ago. If 'still believes' requires the content of the beliefs be the same, then the propositional content of the wife's belief after 30 years is *My husband is a police officer at t*.

Likewise, in the envisaged example outlined in LOST LOVE, Peter's original belief has the propositional content *Barbara loves me at t*, where *t* is some time 10 years ago. If 'still believes' requires the content of the beliefs be the same, then the propositional content of Peter's belief after 10 years is *Barbara loves me at t*.

Finally, in the envisaged example outlined in DEFENSE, the supervisor's original belief has the propositional content *S's dissertation is done at t*, where *t* is the time of the student and her supervisor's exchange. If 'still believes' requires the content of the beliefs to be the same, then the propositional content of the advisor's belief a few weeks later is *If you still think that it's done at t, then we'll talk*.

But it is hardly the case that the wife in DECEIT means that she still believes the same time-indexed proposition after 30 years, *viz*. the proposition *my husband is a police officer at t*, where *t* is some time 30 years ago, that the friend in LOST LOVE means that Peter still believes the proposition *Barbara loves me at t*, where *t* is some time 10 years ago, or that the supervisor in DEFENSE is asking *S* to return if *S* still believes the proposition *S's dissertation is done at t*, where *t* is the time of their exchange. To my mind, such cases raise one of the most pressing kinds of problems for eternalism.

2.4 Arguments from Disagreement

Eternalism holds that present-tensed sentences make implicit reference to the time of speech. 'John is a firefighter,' for example, expresses, relative to a context, the proposition that John is a firefighter at t^*, where t^* is the time of speech. But once we insist that the

contents of our utterances refer to a fixed time, it becomes difficult to see how we can have proper agreements and disagreements over extended time periods. So, successful communication over time must at least sometimes involve temporal contents. Or so I will argue.

The style of argumentation here is similar to the one used by relativists and non-indexical contextualists to refute more general forms of contextualism. This style of argumentation has received its fair share of criticism most recently in Herman Cappelen and John Hawthorne's (2009) *Relativism and Monadic Truth*. This is not the place to engage in the broader debate about whether this form of argumentation can be successfully employed in a refutation of indexical contextualism. Here I will just look at the localized case of eternalism. I reply to Cappelen and Hawthorne's criticisms pertaining to this localized case below.

One way in which arguments from disagreement presented against eternalism differ from arguments from disagreement presented against indexical contextualism more generally is that the former arguments are specifically directed at the claim that all tensed propositions make reference to specific times. But conversations take place over extended periods of time, and most of these conversations are not about specific times in the recent or not so recent past but about some other subject-matter altogether. Specific times may be completely irrelevant to what is discussed. So, it seems that the information that is passed on and that is the subject of discussion in many cases is temporally neutral. It should therefore not come as a surprise if conversations that take place over time become real challenges for eternalism even if they provide no real problem for broader indexical contextualist theories.

To see why the eternalist may have trouble accounting for how information is passed on in ordinary conversations, consider the following exchange:

FIRED FIREFIGHTER
(A and B are talking on the phone. B is standing outside the door of an office where a conversation is taking place between John and his superior).
A: … John is a firefighter.
(Behind closed door the superior is shouting: "you are fired!")
B: I guess you are right. But John is not a firefighter. He was just fired.

The discourse fragment is supposed to sound odd. If you don't have that intuition, the argument does not even get off the ground. However, most people seem to have the intuition that the discourse fragment sounds odd. But let us look now at the predictions yielded by a standard version of eternalism that takes propositions to make reference to a specific time. On such a version of eternalism, A says that John is a firefighter at t_1, and B then replies that A is right but adds that John is not a firefighter at t_2. Notice that there is nothing wrong with the translation I just provided. It doesn't sound odd at all, and for good reasons. If A said that John is a firefighter at t_1, then we should expect B's reply to be acceptable. For it is still true at t_2 that John is a firefighter at t_1.

However, in the envisaged scenario, it would make much more sense for B to have replied: "No, you are wrong. I am standing outside the superior's office, and the superior just told him that he was fired."

Note that this argument, as formulated, does not rest on any intuition about whether A asserts a proposition denied by B. Rather, the argument rests on the oddity of the discourse fragment together with a version of eternalism that takes propositions to make reference to specific times. In other words, if FIRED FIREFIGHTER sounds odd, but the eternalist translation does not, then the eternalist translation is likely mistaken.

Tsompanidis (2013) raises several objections to this type of argument, which I presented in *Transient Truths* (Brogaard, 2012). I will briefly review what I consider his main objection to this type of argument. He argues that the eternalist could turn to interval semantics to account for agreement and disagreement. For example, 'John is a firefighter' might mean 'John is a firefighter *at least up to* and *including* the time of *the entire conversation*.' This type of account may be able to explain what is wrong with dialogues like the one presented in FIRED FIREFIGHTER. As Tsompanidis notes, I *do* consider this kind of reply at length in the book but let me address the specific account he proposes. One major problem for defenders of this type of proposal is to give precise truth-conditions for sentences, given that conversations do not have clear boundaries. A further, related, problem is that the time of the entire conversation cannot always serve as a reference time. Consider the following sentences:

(7)
(a) Mary is falling down from the tree
(b) Afghanistan is at war
(c) I am alive.

If (7a) is uttered during an extended conversation that may continue for hours while Mary is taken to the hospital, the relevant time interval cannot be one that includes the entire conversation. In this case, it may be suggested that the time interval is determined by the duration of the event. However, this suggestion cannot be right. I might utter (7a) because I believe that Mary is falling down from the tree, even though she is not. In that case, there is no event to determine the relevant time interval. While there are many other proposals that could be considered, the sentences in (7a–c) suggest that it will be difficult to give a systematic account of the time intervals that the present tense is supposed to make reference to. Though I agree with Tsompanidis that there are very many points that need to be settled about how language makes reference to time, I think that the problems the eternalist encounters with respect to agreement and disagreement give us a strong reason to prefer temporalism to eternalism.

In their monograph *Relativism and Monadic Truth* Cappelen and Hawthorne provide evidence against disagreement data and argue that the best test for whether an expression is context-sensitive or not is one that gives "center stage to the verbs 'agree' and 'disagree'" (2009, p. 54). The test can be illustrated by means of an example. If A says 'Mary has had enough. She has had three slices of cakes' and B says 'Mary has had enough. She is going to leave her husband,' then we cannot correctly infer 'A and B agree that Mary has had enough.' The oddity of the agreement report is supposed to show that 'had enough' is context sensitive. It has different meanings in different contexts.

The reason the test works as a true test of shared content, Cappelen and Hawthorne say, is that it is hard to hear 'agree' in agreement reports as distributive. Cappelen and Hawthorne then argue that the test shows that propositions are not temporally neutral. Here is one of their examples. John says 'Bill has died' in response to the question 'Why did Bill not show up at the pub last week?' And Janet says 'Bill hasn't died' in answer to the question 'Why did Bill's children not get their inheritance last year?' They conclude that "The claim 'Janet and John disagreed about whether Bill had died' is clearly infelicitous" (2009, p. 98).

However, Cappelen and Hawthorne's test fails. For disagreement to take place it is not sufficient that one speaker denies something that another speaker asserts. Interesting

disagreement requires that there is a time at which two speakers are, or pretend to be, in the same conversational context and are prepared to assign different truth-values to the same content.[1] In the envisaged scenario, John and Janet are not, and do not pretend to be, in the same conversational context. So, they don't disagree in any interesting sense. Hence, the disagreement report is false.

Consider the following modified example: John and Janet are having a dispute about whether Bill has died. John says: 'Bill has died. He didn't show up at the pub last week.' Janet replies: 'No, Bill hasn't died. His children didn't get their inheritance.' Given this conversational context, the disagreement report 'Janet and John disagree about whether Bill has died' comes out true.

Cappelen and Hawthorne argue that 'debated' has the same properties as 'agree' and is equally suitable for testing for context sensitivity (2009, p. 57). Substituting 'debated' for 'agreed,' however, gives us the same results. 'Janet and John debated whether Bill had died' is false in the first case and true in the second.

To further see that the disagreement test fails, consider the following example. John says 'Bill died at 2 p.m., December 11, 2010 EST' in response to his drinking buddy's question 'Why did Bill not show up at the pub last week?' And Janet says 'Bill didn't die at 2 p.m., December 11, 2010 EST' in response to her husband's question 'Why didn't Susan win the bet?' Here the claim 'Janet and John disagreed about/debated whether Bill died at 2 p.m., December 11, 2010 EST' is clearly false, despite the fact that Janet denies what Bill asserts. This becomes even more apparent if we make the innocent move of substituting 'had a disagreement about' for 'disagreed about.' Janet and John did not have a disagreement about anything. But we cannot take that to mean that 'Bill died at 2 p.m., December 11, 2010 EST' has different meanings in different contexts.

2.5 Temporalism and the Problem of Intentionality

The arguments provided above give us some reason to favor temporalism over eternalism but the most compelling argument, in my opinion, turns on the problem of intentionality. The problem of intentionality is that of explaining how a set or a mereological sum of objects and properties comes to represent anything. How does a set of individuals and properties come to have intentional properties? This is a problem that goes back at least to Frege and the early Russell. It is also sometimes misleadingly known as the 'problem of the unity of the proposition.' Recently, several philosophers of language have argued that the problem cannot be cracked if we keep treating sets of individuals and properties as the entities that do the representational work (see, e.g., Soames, 2013; Brogaard, 2014). Although we still need to figure out what '(conscious) representation' means, a first step in the right direction is to realize that intentionality is first and foremost a property of cognitive states. What is called a 'proposition' (or a 'content') is best understood as a kind of generalization based on token cognitive states. We can take propositions to be types of cognitive acts, an act of predicating involved in perceptual states, belief states, agreements, and so on. Propositions thus have representational properties only in a derivative sense. They themselves are generalizations based on token cognitive states. It's the token cognitive states that are the primary bearers of intentional properties.

But, now, not all token mental states represent times. Here is a counter-example to the assumption that all mental states represent times. Mary is pregnant on December 24, 2014, and is expected to give birth on January 15, 2015. But on the morning of December 24,

2014, John and Mary are in a car accident. Mary and the baby are fine. But John is in a coma. Exactly four months later John wakes up and remembers the accident, up to his losing consciousness. He believes it is still December, 2014, and says: 'Where is Mary? She is pregnant.' It is reasonable to think that John really believes that Mary is pregnant. But it just isn't true that John believes on April 24, 2015, that Mary is pregnant on April 24, 2015. He knows human pregnancy cannot take 12 months. He believes Mary is pregnant because he believes it's still December, 2014. So, John's belief that Mary is pregnant cannot plausibly be taken to represent that Mary is pregnant on April 24, 2015. A more plausible suggestion is that John believes in a temporally neutral way that Mary is pregnant. Cases like this give us good reason to think that not all mental states represent times.

If cases like this one aren't sufficiently convincing, we can turn to the case of perception. Perceptual experience does not seem to represent times. You can perceive the same visual scene at 1:01 p.m. and at 1:02 p.m. without the phenomenology of your visual experience having changed one bit. So, the phenomenology of visual experience does not always represent times, which is to say that not all acts of predicating represent a time. But propositions are generalizations from acts of predicating involved in perception, belief, and wishing, and so on. We generalize away differences. As not all acts of predicating represent times, propositions do not in general represent times. But if they do not, then eternalism is false.

3 The Quantifier View versus the Operator View

3.1 Evidence against the Operator View

As we saw above, Kaplan's argument for temporalism rests on the premise that there are tense operators in the English language. This makes the debate about tenses directly relevant to the debate about temporalism versus eternalism. But there are other reasons to consider the two debates logically interconnected. Though temporalism is not articulated as a view about how to treat the tenses in English, on the most natural understanding of temporalism, the debate between temporalism and eternalism is not orthogonal to the debate about how to treat the tenses. Standard versions of eternalism require that the time of speech is a constituent of all propositions. As the time of speech is variable, sentences that express eternal propositions must have a hidden variable in the sentence structure that takes times of speech as its values. This type of sentence structure follows as a natural consequence of a treatment of the tenses as quantifiers. Where t^* is the time of speech, 'John is a firefighter' is of the form 'John is firefighter at t^*,' 'John was a firefighter' is of the form 'there is a time t such that t is earlier than t^*, and John is a firefighter at t,' and 'John will be a firefighter' is of the form 'there is a time t such that t is later than t^*, and John is a firefighter at t.'

Temporalism, by contrast, must treat the tenses as sentential operators, at least given standard semantics. It may be thought that it is possible to combine temporalism with a quantificational account of the tenses. For example, it may be thought that 'John was a firefighter' could be treated as having the following underlying form:

(8) $\exists t(t < t_n$ & John is a firefighter at $t)$,

where t_n is an unarticulated constituent that takes different values across time. If (8) expresses a proposition with an unbound variable, then that proposition will have different

truth-values at different times. The problem with this view is that a content that contains an unbound variable isn't a complete proposition, given standard semantics. In standard semantics, sentences, relative to context, express complete propositions that do not require further satisfaction by context. So, unless we adopt some special semantics, (8) expresses an eternal proposition, *viz*. the proposition that results from substituting the time of speech for t_n. It thus seems that within a fairly standard semantic framework, temporalism is committed to a treatment of the tenses as sentential operators, whereas eternalism is committed to a treatment of the tenses as quantifiers over times or some similar view (e.g., a treatment of the tenses as quantifiers over events or as discourse variables).

The problem for the temporalist is that a wide range of empirical evidence suggests that the tenses function as quantifiers (or perhaps variables) and not as circumstance-shifting operators. I cannot cover the full range of evidence here. But a few illustrative examples are in order. King (2003) offers three main pieces of evidence to motivate a shift from the operator view to the quantifier approach. One consideration against the standard treatment is that it gives us the wrong truth-conditions for sentences with time adverbials. Consider, for instance (King 2003, p. 216; Dowty, 1982, p. 23):

(9) Yesterday, John turned off the stove.

According to King, traditional tense logic would treat (9) as featuring two operators, namely the simple past tense (P), and 'yesterday' (Y). Y shifts the time of speech to some time yesterday, and P shifts the time of speech t^* to some time t such that t is earlier than t^*. Since (9) contains two operators, says King, it should have the following two readings:

(9)
(a) Y(P(John turns off the stove))
(b) P(Y(John turns off the stove)).

(9a) says that the day before some time in the past John turned off the stove, whereas (9b) says that John turned off the stove some time before yesterday. But (9a) and (9b) do not give us the correct readings for (9). (9a) is true just in case John turns off the stove the day preceding some past time, and (9b) is true just in case John turns off the stove at some time past of yesterday. So (9a) and (9b) may both be true if John turned off the stove 10 days ago; but (9) would be false. Thus, a traditional tense logic yields the wrong truth-conditions for sentences like (9). King's quantificational analysis makes the correct predictions. On King's analysis, (9) cashes out to: 'there is a past time t such that t was some time yesterday and John turns off the stove at t.'

A second reason King offers against a treatment of the tenses as sentential operators is that it would make the wrong predictions in cases like the following (2003, p. 217):

(10) Sheila had a party last Friday, and Sam got drunk.

As Barbara Partee (2004) has made vivid, the English tenses can be anaphoric on other tenses in much the same way that pronouns can be anaphoric on quantifiers or terms. The idea is that (10) is similar in important respects to:

(11) Sam took the car yesterday, and Sheila took it today.

In the case of (11), the pronoun 'it' in the second clause is anaphoric on 'the car' in the first sentence. On one theory of unbound anaphora, defended by Stephen Neale (1990) and others, unbound anaphoric pronouns go proxy for definite descriptions recoverable from the antecedent clause. The 'it' in 'Sheila took it today,' for example, goes proxy for the definite description 'the car Sam took yesterday.' Likewise, in (10) the past tense of the first clause picks out a time interval that is supposed to fall within the time interval picked out by 'last Friday.' The past tense of the second clause is anaphoric on the interval picked out by the past tense of the first sentence. The second clause is thus interpreted as meaning that Sam got drunk at Sheila's party last Friday.

The problem for theories that treat the tenses as sentential operators is that if 'last Friday' in (10) and the past tense of the first clause are treated as independent operators, then the second conjunct in (10) receives the implausible interpretation that Sam got drunk at some time in the past, which – if Sam is like most of us – is obviously true. Again, a treatment of the tenses as sentential operators seems to yield the wrong truth-conditions.

According to King, a more debilitating problem for theories that treat the tenses as operators is that they are unable to give a convincing account of Kamp/Vlach sentences such as:

(12)
(a) One day, all persons alive now will be dead
(b) Once all persons alive then would be dead.

The problem that such sentences present is that they have no satisfactory paraphrase using only the resources of traditional tense logic. In traditional tense logic the future tense operator (F), when unembedded, shifts the time of evaluation from the present time to some time in the future. Anything that occurs within the scope of the future tense operator is evaluated with respect to that time, which makes it difficult to translate (12a). (12b) presents a different problem. The problem here is that the past evaluation time is lost when the future evaluation time is introduced. To translate (12a) and (12b), King says, we need to introduce something like Hans Kamp's (1971) doubly indexed N operator, and Frank Vlach's (1973) doubly indexed K operator, which requires a rather complicated and undesirable semantics.

3.2 Complex Tense Operators

In previous work I have responded to the first two pieces of evidence by introducing complex tense operators (see, e.g., Brogaard, 2012, ch. 4). Complex tense operators are complexes of basic tense operators and time adverbials. The time adverbial needn't be explicitly mentioned but may be implicitly assumed in the context.

King argued that one apparent problem with a treatment of the tenses as tense operators is that it gives us the wrong truth-conditions for sentences with time adverbials. This problem goes away, however, if we allow the tenses to interact with time adverbials, as in 'it was the case yesterday.' If we allow complex tense operators in the semantics, we can provide the following paraphrase of (9):

(9*) It was the case yesterday (that John turns off the stove).

'It was the case yesterday' functions as a circumstance-shifting operator that maps *John turns off the stove* to the true iff *John turns off the stove* is true at a past circumstance of evaluation whose time feature belongs to the class of times picked out by 'yesterday.' Of course, English requires that the embedded clause in (9*) occurs in the past tense. So, in ordinary English, (9) should be paraphrased as 'It was the case yesterday that John turned off the stove.' On the relevant reading, the past tense of the embedded clause is vacuous. There is also an alternative reading where the past tense of the embedded clause is not vacuous. For example, if John turned off the stove the day before yesterday, we can truthfully utter the sentence 'It was the case yesterday that John turned off the stove two days ago.'

Related considerations help to address cases where the tense of one clause is anaphoric on the tense of a preceding clause, repeated from above:

(10) Sheila had a party last Friday, and Sam got drunk.

Here the past tense of the first sentence picks out a time interval that falls within the time interval picked out by 'last Friday.' The past tense of the second clause is anaphoric on the interval picked out by the past tense of the first clause. Two key cases cited as evidence against the operator account thus turn out not to present a threat to the account.

3.3 Montague Grammar

That still leaves us with Kamp/Vlach sentences. Such sentences, it turns out, can be dealt with satisfactorily in Montague grammar (PTQ – proper treatment of quantification in ordinary English; Montague, 1973), which is a tense logic. Consider:

(13) A colleague of mine who was a child prodigy got her PhD from Harvard.

It is tempting to think that we can get the following reading with nested clauses: '$\exists x$(colleague x & P(get PhD x & P(prodigy x))).' On this reading, (13) says that there is someone who is currently at colleague who got her PhD from Harvard at some point in the past and who was a child prodigy before that. Unfortunately, this reading cannot be yielded compositionally.

Compositionality requires that a meaning is yielded for the noun phrase 'A colleague of mine who was a child prodigy,' and that this meaning is then combined with the meaning of 'got a PhD from Harvard.' So, given a compositional interpretation of English syntax (with the exception that noun phrases can scope out), it is not possible for the past tense in 'was a child prodigy' to have wider scope than 'colleague of mine.'

The reason that a meaning is yielded for the whole noun phrase 'A colleague of mine who was a child prodigy' is that within the whole noun phrase 'A colleague of mine who was a child prodigy,' the relative clause is a self-contained syntactic constituent. This constituent has the syntax of a full sentence except that it lacks a noun phrase. Instead of a noun phrase it has a variable that is bound by the noun phrase. So, the relative clause is of the form 'x who was a child prodigy,' where the variable 'x' is bound by 'a colleague of mine.' The compositional structure of 'A colleague of mine who was a child prodigy' is as follows. 'A colleague of mine who was a child prodigy' is composed of the indefinite article 'a' and the noun phrase 'colleague of mine who was a child prodigy.' The latter is composed of 'colleague of mine' and 'who was a child prodigy,' which in turn is composed of 'who' and 'x

was a child prodigy.' Finally, '*x* was a child prodigy' is composed of the past tense morpheme and '*x* is a child prodigy.'

As PTQ observes compositionality (with the exception of the 'quantifying in' rule), it yields the following readings for (13):

(13) A colleague of mine who was a child prodigy got her PhD from Harvard.
(13a) $\exists x$(colleague x & P(prodigy x) & P(get PhD x))
(13b) P($\exists x$(colleague x & P(prodigy x) & get PhD x))

(13a) translates as 'some colleague is such that it was the case that she is a prodigy, and it was the case that she gets her PhD' and (13b) translates as 'it was the case that some colleague, who was a prodigy, gets her PhD.' In (13a) there is quantifying in: the whole noun phrase 'A colleague of mine who was a child prodigy' has wider scope than the main clause, in (13b) there is no quantifying in. So the whole noun phrase 'A colleague of mine who was a child prodigy' has scope under the tense of the main clause.

Dan Zeman (2013) has responded to these sorts of strategies that the temporalist might offer to preserve a fairly traditional tense logic that while the temporalist no doubt can come up with an operator account of the tenses that can accommodate most, if not all, of the phenomena that normally are cited in support of the quantificational account, true supporters of temporalism might want "positive, decisive arguments for the view that tenses are to be interpreted as circumstance-shifting sentential operators, rather than, say, quantifiers over temporal variables verbs come endowed with" (Zeman, 2013, p. 325).

I agree with Zeman that there are very few empirical data concerning the semantics of tense that cannot be accommodated by both operator accounts and quantificational theories of the tenses (as well as many other theories of the tenses). I also agree that debates about the tenses are not going to settle the debate about temporalism. Rather, what is ultimately going to settle the debate between temporalism and eternalism is an argument that is independent of how we treat the tenses in the English language. If the debate can be settled in favor of temporalism, as I have argued above, that gives us good reasons for revisiting an operator theory of the tenses.

4 From Philosophy of Language to Metaphysics

4.1 The Incompatibility of Presentism and Semantic Eternalism

I will now turn to the important question of whether the debate about temporalism versus eternalism has any bearing on the debate about presentism versus metaphysical eternalism. In *Transient Truths* (2012) I argue that the answer is 'yes.' Semantic eternalism, for example, appears to be inconsistent with presentism, a particular version of the A-theory. The argument is this. Presentism holds that only present things exist. But according to the standard version of semantic eternalism, all propositions include a timestamp (e.g., the sentence 'Mary is hungry' may express the proposition that Mary is hungry at 2:05 p.m. on October 1, 2013 EST). Most of these timestamps are past and future times. So, if presentism is true, then the vast majority of these propositions do not exist. The presentist could construe times as ersatz times (sets of propositions) (Brogaard, 2013). But on pain of circularity, this requires granting that there are temporal propositions (without a timestamp). So, presentism is at odds with semantic eternalism.

Giuliano Torrengo (2013) has replied that the book's argument doesn't work, because temporalism is consistent with there being some eternal propositions, for example, the propositions that there are wholly past objects and that I am giving a talk at Stanford University on May 15. Yet, Torrengo argues, "it is the thesis that *some* eternal propositions exist that is at odds with presentism" (Torrengo, 2013, p. 316). This is a nice point. However, I disagree with Torrengo that presentism is at odds with the thesis that there are *some* eternal propositions. As Torrengo himself points out, it is the stronger view that there are *no* temporal propositions (i.e., semantic eternalism) that prevents the presentists from construing times as ersatz times. The weaker view defended in *Transient Truths* (i.e., temporalism) leaves us with all the resources (i.e., temporal propositions) needed to construe times as sets of propositions non-circularly.

That said, Torrengo is perfectly right that if presentism is true, then the temporalist cannot accept all of the eternal propositions ordinary language appears to commit us to. Some temporalists (myself included) argue (while bracketing metaphysical issues) that there are eternal propositions that make explicit reference to times, for instance, the proposition that I am giving a talk at Stanford on May 15. If presentism is true, then that proposition does not currently exist. Presentists must, therefore, reject the existence of these kinds of propositions. (They can, of course, accept the existence of metaphysical propositions such as *there are wholly past objects*, as these types of propositions do not have times as constituents.) The thought that sentences, such as 'I am giving a talk at Stanford on May 15,' do not express a proposition at all and therefore are false is not entirely unmotivated. It could be argued that while an utterance of the sentence 'I am giving a talk at Stanford on May 15' may seem true, this kind of speech is, in fact, idiomatic much like 'the sun is rising.' Idiomatic speech is literally false (or untrue) but conveys something true.

4.2 The Incompatibility of Metaphysical Eternalism and the Quantifier Theory of the Tenses

How we treat the tenses also seems to have metaphysical implications. There is some reason to think that if metaphysical eternalists adopt the quantifier account of the tenses (that is, the semantic eternalist's common account of the tenses), then they will have difficulties making certain metaphysical claims (Brogaard, 2012, ch. 7). The gist of the argument runs as follows. The metaphysical eternalist wants to say that past and future objects exist *simpliciter*. Consider:

(14) Socrates exists.

According to the metaphysical eternalist, Socrates existed in the past but does not presently exist. So, the metaphysical eternalist holds that (14) is true on one reading but false on another. Now combine metaphysical eternalism with the quantifier account of the tenses. On the quantifier account, all propositions are indexed to a time. So, where t^* is the time of speech, (14) is equivalent to the proposition expressed by (15):

(15) Socrates exists at t^*.

But here is the problem. If (15) specifies a false proposition expressed by (14), then what is the nature of the true proposition expressed by (14), according to the metaphysical eternalist?

Torrengo has replied to this sort of argument that "once we accept the distinction between a temporally restricted and a temporally unrestricted reading of quantification (and something analogous for predication), the worry is spurious" (Torrengo, 2013, p. 318). However, this misses the point of the argument. The argument is that if the metaphysical eternalist accepts a quantificational account of the tenses, then she cannot account for the unrestricted reading of (14). (14) can, of course, be read as follows (as Torrengo suggests):

(16) $\exists x(x = \text{Socrates})$, where the domain of values is temporally unrestricted.
(17) $\exists x(x = \text{Socrates})$, where the domain of values is restricted to the present.

According to the metaphysical eternalist, (16) then is true and (17) false. However, this proposal is compatible with a version of temporalism that utilizes quantifier restriction. My own proposal was similar. On the view I prefer, (14) has a reading that determines a function from worlds to extensions and another reading that determines a function from world-time pairs to extensions. The first reading is the 'unrestricted' reading, whereas the second is the 'restricted' reading.

Notice, however, that neither of these proposals utilizes a quantifier account of the tenses. In fact, they are inconsistent with the standard version of semantic eternalism, which requires that all propositions are indexed to a time. And that was just the point of the argument, which was not an argument against metaphysical eternalism but one in favor of temporalism (on the assumption that metaphysical eternalism is true).

A second worry that Torrengo raises is that the presentist cannot claim that (14) is false, on an unrestricted reading. The reason for this, he says, is that I hold that an eternal proposition such as *Socrates exists* is "evaluable as true or false *simpliciter* only in a context in which either Socrates is an instantaneous object, or Socrates always (or never) exists" (2013, p. 318). However, this is not my view. What I said was:

> I think that one *could* use 'John has a straight shape' to mean the eternal proposition that John has a straight shape. But such an eternal claim is truth-evaluable at a world *w* only if (i) John is an instantaneous object at *w*, (ii) John always has a straight shape at *w*, (iii) John never has a straight shape at *w*, or (iv) Lewis is right that the *eternal* proposition *John has a straight shape* is true at *w* iff John has a temporal part that has a straight shape. (Brogaard, 2012, p. 150)

I made this remark in the context of discussing Lewis's problem of temporary intrinsics. The reason 'John has a straight shape' cannot be evaluated except under these conditions is that if John sometimes has a straight shape and sometimes has a bent shape, then relative to the world as a whole the proposition is neither true nor false (or both true and false). The same point does not apply to the proposition that Socrates exists (as existence does not come and go).

A third concern that Torrengo raises also concerns the unrestricted reading of sentences like (14). He argues that my view implies that the unrestricted readings of sentences are never true for the presentist, not even when the entity in question is present. This has the consequence, he says, that "the presentist and the eternalist necessarily disagree on what *presently* exists" (Torrengo, 2013, p. 320), which seems odd.

I agree with Torrengo that that would be odd. However, I don't think the temporalist is committed to this view. Consider:

(18) Obama exists.

If presentism is true, then (18) is true when read restrictedly and unrestrictedly. On the restricted reading, 'Obama' determines a function from world-time pairs to extensions. Since the extension is non-empty, (18) is true on this reading. On the unrestricted reading, 'Obama' determines a function from worlds to extensions. Since this extension is also non-empty, (18) is true on this reading. Torrengo thinks I cannot say this, because I argue that on the unrestricted reading, (18) entails that it will be the case that Obama exists. However, even if we bracket Obama's future existence, there is no problem here, because this kind of tensed sentence is innocuous. It is the result of affixing a tense operator to a sentence given an unrestricted reading. But when tense operators are affixed to an operand sentence that expresses an eternal proposition, the tense operators will be redundant (Brogaard, 2012, p. 150). So, the presentist can agree with the metaphysical eternalist that (18) is true on both its restricted and its unrestricted reading.

Torrengo is right that if the presentist holds that Obama is not fully present but is unfolding in time, then it would seem that she should reject (18). After all, if only some of Obama's parts exist, how could (18) be literally true? I think this is a genuine puzzle but not one that is specifically about the unrestricted reading of (18). It appears to be equally problematic on the restricted (ordinary) reading of (18). The puzzle is not a consequence of accepting presentism or temporalism. Anyone who holds that ordinary material objects are four-dimensional space–time worms needs a way to talk about the properties the present parts instantiate. This is a familiar issue from the metaphysical literature (see, e.g., Sider, 2001). It is true that Obama is speaking even if it's only his present part that is speaking. Yet how can this be if he is extended four-dimensionally? One standard reply is that proper names ordinarily refer only to stages of objects. Whether this is the best reply to the worry is not something I can address here. But let me point out that most three-dimensionalists who take ordinary material objects to endure are faced with a version of this problem. It is commonly agreed upon that events perdure: they have temporal parts located at different times. Yet even if a soccer match takes a considerable amount of time, it can nonetheless still be true to say that you are currently watching one. So, the problem of how to correctly predicate properties of four-dimensional entities may arise regardless of one's particular view of how ordinary material objects persist through time.

4.3 The Passage View and Monadic Truth

As argued above, presentism necessitates the existence of temporal propositions. In reply to my arguments John Hawthorne (2013) has argued that the need for the index of evaluation to contain a time parameter goes away if one accepts what we might call 'the passage view.' Unlike presentism, which holds that only present things exist, the passage view holds that future and past things exist but that they have a different status compared to presently existing things. One option is to treat past and future objects as abstract and present objects as concrete. Another option is to say that only present events are happening, which is Hawthorne's preferred version of the view. The details of this position need not concern us here. What matters is that the passage view, like presentism, can hold that there is something special about the present moment, but unlike presentism it does not need to deny the existence of past and future entities. Because presentism and the passage view take the present to be special, Hawthorne argues, presentists and defenders of the passage view need not relativize contents to times or stipulate that there are times in the index of evaluation.

According to him, we can simply take propositions to bear truth-values *simpliciter*. This renders temporalism obsolete. However, I think this argument is unsound. I have already argued that the presentist is committed to the existence of temporal propositions. A different argument is required to show that the defender of the passage view is also committed to the existence of temporal propositions and indices of evaluation that include times. Consider the sentence:

(19) It was the case that (there is a time t, and dinosaurs exist at t).

On the passage view, the domain of objects remains constant across time, as no object comes into existence or ceases to exist. An individual might become abstract after having been concrete but it doesn't go out of existence. But when we have a constant domain of individuals, then the Barcan and converse Barcan formulas are true. Hence, (19) entails:

(20) There is an x such that it was the case that x is a time and dinosaurs exist at x.

So, there are two times: the present time and x (a past abstract time). But if there are two times, then it seems that we need to relativize to times, even on the passage view.

5 Conclusion

Two of the main debates about tense and time in philosophy of language concern the eternalism–temporalism dispute and the semantics of the tenses. I have argued here that while these debates are far from settled, there are currently more significant pointers to temporalism than eternalism as an adequate account of propositions. This may impact our choice of semantics of the tenses in the English language. Temporalism seems to require a more traditional tense logic that treats the tenses as circumstance-shifting operators.[2]

Notes

1 If you disagree with Aristotle that, say, love is a union, then you are not in the same conversational context, but you pretend to be by entering a dialogue with the asserter of the claims you find problematic.

2 For helpful comments on a previous version of this paper I am grateful to Alex Miller and Bob Hale.

References

Brogaard, B. 2012. *Transient Truths*. New York: Oxford University Press.

Brogaard, B. 2013. "Presentism, primitivism and cross-temporal relations: lessons from holistic ersatzism and dynamic semantics." In *New Papers on the Present: Focus on Presentism*, edited by R. Ciuni, K. Miller, and G. Torrengo, pp. 253–280. Munich: Philosophia Verlag.

Brogaard, B. 2014. "An empirically informed cognitive theory of propositions." *Canadian Journal of Philosophy*, 43(5): 534–557.

Cappelen, H., and J. Hawthorne. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Dowty, D. 1982. "Tenses, time adverbs, and compositional semantic theory." *Linguistics and Philosophy*, 5(1): 23–55.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Hawthorne, J. 2013. "Comments on *Transient Truths*." Pacific Division Meeting of the APA, March 30, 2013.

Kamp, H. 1971. "Formal properties of 'now.'" *Theoria*, 37(3): 227–273.

Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. New York: Oxford University Press.

King, J. 2003. "Tense, modality, and semantic values." *Philosophical Perspectives*, 17: 195–246.

Montague, R. 1973. "The proper treatment of quantification in ordinary English." In *Approaches to Natural Language*, edited by K. Hintikka, J. Moravcsik, and P. Suppes, pp. 221–242. Dordrecht, Netherlands: Reidel.

Neale, S. 1990. "Descriptive pronouns and donkey anaphora." *Journal of Philosophy*, 87(3): 113–150.

Partee, B. 2004. "Some structural analogies between tenses and pronouns in English." *Journal of Philosophy*, 70(18): 601–609. Reprinted in *Compositionality in Formal Semantics. Selected Papers by Barbara H. Partee*, pp. 50–58. Oxford: Blackwell, 2004.

Richard, M. 1981. "Temporalism and eternalism." *Philosophical Studies*, 39(1): 1–13.

Salmon, N. 1989. "Tense and singular propositions." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 331–392. New York: Oxford University Press.

Sider, T. 2001. *Four-Dimensionalism: An Ontology of Persistence and Time*. Oxford: Clarendon Press.

Soames, S. 2013. "A cognitive theory of propositions." In *New Thinking about Propositions*, edited by J. King, S. Soames, and J. Speaks. Oxford: Oxford University Press.

Stanley, J. 1997a. "Names and rigid designation." In *A Companion to the Philosophy of Language*, edited by B. Hale and C. Wright, pp. 555–585. Oxford: Blackwell.

Stanley, J. 1997b. "Rigidity and content." In *Logic, Language and Reality: Essays in Honor of Michael Dummett*, edited by R. Heck, pp. 131–156. Oxford: Oxford University Press.

Torrengo, G. 2013. "Propositions and the metaphysics of time." *Disputatio*, 5(37): 315–321.

Tsompanidis, V. 2013. "On two arguments for temporally neutral propositions." *Disputatio*, 5(37): 329–337.

Vlach, F. 1973. "'Now' and 'Then.' A Formal Study in the Logic of Tense Anaphora." PhD diss., University of California Los Angeles.

Weber, C. 2013. "Propositions and Centered Content." PhD diss., Australia National University.

Zeman, D. 2013. "Temporalism and composite tense operators." *Disputatio*, 5(37): 323–328.

Further Reading

Frege, G. 1952. "On sense and nomination." In *Translations from the Philosophical Writings*, translated by P. Geach and M. Black. Oxford: Blackwell.

Ludlow, P. 1999. *Semantics, Tense, and Time: An Essay in the Metaphysics of Natural Language*. Cambridge, MA: MIT Press.

30

Relativism

PATRICK SHIRREFF AND BRIAN WEATHERSON

Relativism, in the sense we're interested in here, is the view that the truth of a sentence is relative both to a context of utterance and to a context of assessment. That the truth of a sentence is relative to a context of utterance is uncontroversial in contemporary semantics. If the authors of this chapter were to both utter the sentence "I'm Canadian," then one of us would say something true, and the other something false. And that's because the truth of the utterance "I'm Canadian" is sensitive to a feature of the context of utterance, namely the identity of the speaker. And that in turn is explained by the fact that the proposition expressed by that sentence is different in different contexts. Relativists deny that this simple story can explain all the ways in which context affects language.

A core motivation for relativism comes from looking at the problems with other views. So let's start by thinking about what might be happening when a speaker says, "This is tasty." (Call this utterance, which we'll come back to a bit, U, and its speaker S.) The demonstrative 'this' is context-sensitive, but let's assume it is clear what is being referred to, and think about what contribution the predicate is making. There are two natural proposals that are simple, easy to fit into familiar frameworks, and most likely wrong.

First, the predicate might pick out an objective property of tastiness, and when S utters U, she ascribes this property to a thing. This objective view has a number of problems. First, there is a metaphysical problem: Just what is this objective tastiness? Second, there is an epistemological problem. Typically, a speaker like S could be prepared to utter U after taking one bite of the thing. But for most plausible answers to the first question, it is unclear how they could know that the object had that property. The two problems reinforce each other; the more plausible answers to the metaphysical question make the property sensitive to very abtruse conditions, such as how idealized observers would react to ingesting the substance. But that makes it even more unlikely that typical utterances of U satisfy the epistemological requirement. (The points we're making here are well known, but our presentation owes a lot to Lasersohn, 2005.)

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Second, the predicate might be context-sensitive. For concreteness, let's focus on a very simple contextualist theory of 'tasty.' An utterance of U is true, relative to a context *c*, iff the speaker in *c* likes the taste of the referent of 'this' in *c*. So "This is tasty" and "I like the taste of this" express something very close to the same proposition. (At least they express propositions that are necessarily materially equivalent, but the relation between the two is probably closer than that.) Again there are two problems, both of them to do with the different dynamics of "This is tasty" and "I like the taste of this." A hearer H, who does not like the taste of this substance, could more readily disagree with "This is tasty" than with "I like the taste of this." If H doesn't like its taste, he'll nevertheless concede that S's utterance of "I like the taste of this" is true, but won't concede that "This is tasty" is true. And a similar phenomenon occurs when S herself changes her mind. If later she dislikes the taste of the substance, she will be disposed to retract her utterance of "This is tasty," but not of "I like the taste of this." So both disagreement and retraction data pose problems for the contextualist. (On disagreement, see especially Tamina Stephenson, 2007. On retraction, and most everything else we'll talk about in this paper, see John MacFarlane, 2014.)

If the objectivist and contextualist solutions fail, the relativist suggests that they have a useful alternative. Say that truth is doubly relativized, both to contexts of utterance and to contexts of assessment. So the utterance U, or perhaps just the proposition expressed by it, is true relative to a context of assessment c_a and context of utterance c_u iff the agent at c_a likes the taste of the denotation of 'this' in c_u. So now the truth of the utterance is relative both to the context of utterance and to the context of assessment. This solves the metaphysical problem; talk about what a person likes is unproblematic. It solves at least a version of the epistemological problem; a speaker knows what they like, so can make utterances that are true in their context. It solves a version of the disagreement problem; if H doesn't like the taste of the stuff, he can truly say that what S uttered is not true, since it isn't true in his context. And it explains the retraction data, since once S changes her taste, what she uttered is no longer true relative to her new context, and hence should be retracted.

As everywhere else in philosophy, arguments and claims can be and are contested. In the quick case for relativism we've described so far, there are at least four points where one could readily disagree.

1. Whether objectivism is really vulnerable to the combination of the metaphysical and the epistemological arguments.
2. Whether this version of contextualism is vulnerable to the disagreement and retraction arguments, and if so, whether these problems can be avoided by a more sophisticated contextualist theory.
3. Whether relativism really does avoid the four problems posed for the other theories.
4. Whether there are other theories that also avoid the problems, without running into the problems facing relativism or problems of their own.

In this chapter, we'll focus on the last three points, since they are more widely discussed in the literature than the first one. And indeed, the last three points have been extremely actively debated in recent years. We won't try to take sides in those debates, though we will note that at no point is the final picture nearly as rosy for the relativist as this initial sketch may suggest.

We've so far focused on one particular predicate, 'tasty,' though the points we've been making generalize to most predicates of personal taste. But this is far from the only area

where relativist theories have been proposed. There has been a lot of discussion of relativist theories of epistemic modals. The arguments here are fairly similar to the arguments about predicates of personal taste (though see the discussion below about syntax and control). And MacFarlane has argued for relativist treatments of future contingents, and knowledge ascriptions. We will not spend much time on those constructions explicitly, but the issues they raise are broadly similar to the issues raised by predicates of personal taste, and by epistemic modals.

We will largely bypass two quite different areas where relativist theories have been proposed. Mark Richard (2008) has argued for a relativist treatment of comparative adjectives. And Brian Weatherson (2011) argued for relativist treatments of certain areas of discourse where common assumptions about the area are false and a relativist treatment might be the most charitable fix. As interesting as we find these proposals, they haven't occasioned nearly as much discussion as the proposals discussed above, and so we'll set them aside.

What we will do is first clarify some of the many ways in which the term 'relativism' has been used in recent debates, then review some technical material about indices and contexts that is essential for understanding some relativist views, then look at the main line of recent debate concerning relativism, the one centered on issues about retraction and disagreement, and finally look at some syntactic evidence for relativism.

1 Varieties of Relativism

As we noted in the very first paragraph, it is uncontroversial that sentence truth is sensitive to the context of utterance. It is extremely contentious just how many such sentences are sensitive to the context of utterance. (See, for instance, DeRose, 2009; Harman and Thomson, 1996; Cappelen and Lepore, 2005, for some of the disputes.) But that there is some sensitivity here is uncontroversial. The view that a sentence-type's truth is sensitive to its context of utterance is sometimes called "indexical relativism" (Kölbel, 2004; Einheuser, 2008). For some purposes this is a useful name. In particular, the view that the content of moral predicates, and hence the truth-value of ascriptions of moral predicates, is sensitive to the context of utterance does seem like a form of moral relativism, as that term is standardly understood. It is, for instance, the view that Harman (1975) defends in a self-proclaimed defense of moral relativism. (That this was traditionally known as relativism is stressed in Cappelen and Huvenes, 2014, who draw some interesting conclusions from this fact.) But for present purposes we want to clearly exclude those views, and focus on much more radical proposals that have been recently made.

There are two (overlapping) families of views that have been called relativist, and which we will be concentrating on for the bulk of this chapter. They are:

- Relativism about propositional truth – Whether a proposition is true is not an absolute fact. Propositions can be true relative to some contexts of assessment, and false relative to others.
- Relativism about utterance truth – Whether an utterance is true is not an absolute fact. Utterances can be true relative to some contexts of assessment, and false relative to others.

It isn't always clear whether a particular author, in defending relativism, is primarily defending the first or second of these claims. But it is natural to interpret most relativists as defending relativism about propositional truth. This is most clearly true for Kölbel (2002) and Egan (2007), but there are few relativists that it is hard to interpret as taking this to be the primary focus. The one big exception, however, is the most prominent defender of relativism in the contemporary literature, John MacFarlane (2014). He takes relativism about utterance truth to be the only genuine form of relativism, though he also endorses relativism about propositional truth. We will stay neutral when it comes to the dispute over what is *really* a relativist view and instead we'll just describe a pair of views that are relativist in one of these senses but not the other, to explain how the two senses come apart.

First, let's think about how to make propositional truth, but not utterance truth, relative. Assume that Suzy actually swims, but that she might not have. Then, on a natural way of thinking about modality, the proposition *Suzy swims* will be true relative to the actual world, but false relative to any possible world in which she does not. So propositional truth is relative, with the relativity being to possible worlds.

Cappelen and Hawthorne (2009) argue that the reasoning of the previous paragraph misunderstands the nature of truth and modality. The proposition *Suzy swims* is, they say, simply true. It might have been false. But that doesn't mean it is merely true relative to some world or other; it just means that it might have had a different property than it actually has. Schaffer (2012) argues that the things we say and think, that is, propositions, typically have worlds in their content, so what we express by the sentence "Suzy swims" is the proposition *Suzy swims at @*, and that proposition has a non-relative truth-value.

Even if all that is wrong, and the contents of sentences like "Suzy swims" are propositions that are true in some worlds, false in others, there is a reason not to call this 'relativism.' After all, other worlds don't exist, or at least don't exist in the way ours does. So it isn't true that the proposition expressed by "Suzy swims" is true relative to something and false relative to something else; there isn't a something else for it to be false relative to.

Taking the last point on board, a serious relativism about propositional truth should say that there is a proposition p, and two relata a and b such that p is true relative to a, and false relative to b. One way to do that, as seen in for example Kaplan (1989), is to make truth relative to times, and believe that other times exist. (The contrast here is with the eternalist view of propositions defended by Evans, 1985.) But the radical view that has been central to many recent versions of relativism has been to say that truth is relative to world, time, individual triples. So there is a proposition, for example, that is true relative to a triple $\langle w, t, a \rangle$ iff in w, the individual a disapproves of murder at t. Perhaps, on a relativist framework, this just is the proposition that murder is wrong. The picture of propositions here owes something to the discussion of *de se* beliefs in Lewis (1979), since Lewis thought the contents of such beliefs were sets of such triples. But Lewis put this forward as a theory about mental content; the move to extending it to linguistic content is more recent.

On this view, the contents of utterances will be propositions that can be true relative to one individual (in a world, at a time) and false relative to another. But the utterances themselves will naturally be thought to have absolute truth-value. Think back to the view that propositional truth is just relative to worlds. Then an utterance of "Suzy swims" will, intuitively, be true iff Suzy swims in the world in which the utterance is made. When we think about a counterfactual utterance of "Suzy swims" in a world where she does not, we don't think the utterance is true just because its content is actually true. So although this is a form of relativism about propositional truth, it is not yet a form of relativism about

utterance truth. MacFarlane, who takes utterance relativism to be central, calls this view 'non-indexical contextualism,' with the name meant to highlight that it isn't, by his lights, genuinely relativist.

Now turn to relativism about utterance truth. A simple way to implement this is to say that the propositional content of an utterance is relative to an assessor. So consider a very simple, and surely false, view about 'you' in English. It holds that the content of an utterance of 'you,' relative to a context of assessment, is the agent of that context. And it says that the utterance is true at a context of assessment iff the proposition it expresses relative to that context is (absolutely) true. So an utterance of "You swim" will express a different proposition relative to contexts with different agents. And since some of those propositions will be true, and some false, we have a version of relativism about utterance truth. But this view is consistent with the view that propositional truth is absolute, not relative even to worlds.

So the two versions of relativism are dissociable. But it is also possible to hold them simultaneously, as MacFarlane (2014) does. To set up MacFarlane's view, it is helpful to review the framework against which it was developed.

2 Index, Context, and Content

The existence of indexical terms, like 'I,' means that we can't give a theory for truth-conditions of sentences as such. The sentence-type *I am in Ann Arbor now* doesn't have truth-conditions; only occurrences of this sentence do. So a natural move is to build up a theory that assigns to each term a function from contexts to contents, and use that to provide a theory of which contexts a sentence is true in. It turns out that there are reasons to endorse more complications than this. In particular, sentences like (1) show the need for relativizing truth-conditions relative to both a context and an *index*.

(1) It might have been the case that my actual parents never met.

That's true. But think of how we might naturally provide truth-conditions for it. First, we define the context-relative truth-conditions for "My actual parents never met." The context will provide a world for 'actually' (i.e., the actual world), and an agent for 'my' (i.e., the speaker). Then, we say that the modal operator shifts the world of evaluation. Intuitively, *Might p* is true iff *p* is true in some possible world. So we need some way to see if the content of "My actual parents never met" is true in some other world. But that's hard to do, since (given origin essentialism), there will be no context in which "My actual parents never met" can be truly uttered, so its content will be the function that maps every context into false. And that's the same content as, for example, "Two plus two is five," and so it can't be possibly true.

There are actually two related problems here. One is that while we need to look around the other worlds to see whether some thing is true, the nature of that thing is fixed by the actual context. So if the sentence is uttered by Sasha Obama in this world, the content is something like *It might have been the case that Barack Obama and Michelle Obama never met*. We need to fix the values of the contextually sensitive terms ('my' and 'actual') even when they appear inside operators. The second problem is that something as coarse-grained as a function is very hard to 'shift' in just one respect.

There is a well-known solution to this, developed primarily by Hans Kamp (1971), then put to important philosophical use by David Lewis (1980) and David Kaplan (1989). Say

that our semantic theory will assign a function from both contexts and indices to terms. Ultimately, each sentence will get as its semantic value a function from contexts and indices to truth-values. Lewis describes the distinction between contexts and indices thus,

> A context is a location – time, place, and possible world – where a sentence is said. It has countless features, determined by the character of the location. An index is an n-tuple of features of context, but not necessarily features that go together in any possible context. Thus an index might consist of a speaker, a time before his birth, a world where he never lived at all, and so on. (Lewis, 1980, p. 79)

Indices are structured, and so they can be 'improper,' meaning that it might consist of things that don't go together in the actual world or any possible world. If the index contains just worlds and individuals, the index could be $\langle w, \text{Sasha Obama}\rangle$, even if Sasha doesn't exist in w. Such an index will play a central role in making (1), as uttered by Sasha Obama, true. A context, on the other hand, will by its nature be proper; contexts will go together in the actual world or some possible world. They will pick out a world, and a time in that world, and a person existing at that time in that world, and so on.

Indices are, unlike contexts, 'shiftable.' There can be sentential operators that say the truth-value of the whole sentence is given by looking at the truth-value of the embedded sentence at some different index. Assume, for example, that there is a world parameter in the index, so the index is $\langle w, x\rangle$, where x contains everything else in the index. Then *Might* p will be true relative to context c and index $\langle w, x\rangle$ just in case for some w', p is true relative to context c and index $\langle w', x\rangle$. Note that in this definition, we do *not* shift c; so if c determines that the referent of 'I' is Sasha Obama, that stays even if we 'shift' the index to a world where Sasha does not exist.

The separation between context and index was developed to solve some technical problems, but it can do some philosophical work. The theory associates each sentence with a function from contexts and indices to truth-values. Equivalently, it associates sentence-context pairs with functions from indices to truth-values. So it is natural to say that the content of an utterance is (or is at least intimately connected to) the function associated with the pair consisting of the sentence uttered and the context it was uttered in. If the index is simply a world, then the contents will be functions from worlds to truth-values, as defended by classical contextualists such as Robert Stalnaker (1984). David Lewis (1980) objected that assigning contents at just this stage was arbitrary; why not associate contents with functions from context to truth-values instead? François Recanati (2007) responds well to Lewis's arguments, although his defense requires making indices much more complex. In particular, he includes parameters for times and individuals in the index, leading him to support relativism about propositional truth.

With this background, we can more easily describe two distinctive features of MacFarlane's views. First, he makes indices very complicated indeed. The index might include, among other things, an information set (for interpreting epistemic and deontic modals), a standard of taste (for interpreting predicates of personal taste), relevant alternatives (for interpreting knowledge ascriptions), a time (for handling future contingents), and perhaps many more things. For simplicity, we'll call these things collectively a *perspective*, and say that indices are world-perspective pairs. (Though note that our terminology here requires perspectives to be structured entities, and for them to be potentially improper.) If

the propositional content of an utterance is (intimately connected to) a function from indices to truth-values, then the one proposition will be true relative to one perspective and false relative to another.

So far we only have relativism about propositional truth, or what MacFarlane calls non-indexical contextualism. MacFarlane argues for a second revision to contextualist orthodoxy. Assume that we have a particular utterance U of a sentence S in context of utterance c_u, and that utterance is being assessed in context c_a. And assume that $w(c)$ is the world of context c, and $p(c)$ is the perspective of context c. Sentence S is associated with a function f_S from context-index pairs, that is, context-world-perspective triples, to truth-values. The relativist about propositional truth, but not utterance truth, says that U is true iff $f_S(c_u, w(c_u), p(c_u)) = TRUE$. MacFarlane's relativist[1] says that U is true iff $f_S(c_u, w(c_u), p(c_a)) = TRUE$.[2] Crucially, we assess the utterance itself, and not just its content, by the perspective of the assessor and not from the perspective from which it was uttered. MacFarlane argues that this move allows a better understanding of disagreement and retraction, which were central to the phenomena that motivated relativism. So let's turn to how well relativism explains the phenomena it was designed to explain.

3 Retraction and Disagreement

3.1 Retraction

Contextualists have a difficult time explaining why we retract earlier claims involving predicates of personal taste or epistemically modal assertions when our perspective has changed in the intervening time. Consider this example:

> Kim (age 8): Lunchables are delicious.
> Kim (age 27 being reminded of previous assertion): I take it back/I was wrong/what I said was false. Lunchables aren't delicious.

When Kim retracts her earlier assertion, it is natural to use any of the three forms of retraction listed here. It isn't natural to say either of the following things:

- Lunchables were delicious back then, but they aren't delicious any more.
- When I said that back then, I only meant that they were delicious to me back then.

There is a contrast with how clearly context-sensitive terms like 'here' are used. It isn't natural to use any of the three forms of retraction we attribute to Kim below.

> Kim (in a café): It is pleasant here.
> Kim (in an oil refinery, reminded of previous assertion): I take it back/I was wrong/what I said was false. It isn't pleasant here.

On the other hand, it is natural to say things like:

- It was pleasant where we were, but it isn't pleasant where we are now.
- When I said that back then, I only meant it was pleasant where we were.

Predicates of personal taste, like 'delicious,' don't behave like explicitly context-sensitive expressions like 'pleasant here.' We see the same pattern with epistemic modals.

HAKEEM: It might be snowing outside.
FABIAN: No/that's wrong/that's false, it can't be snowing. I just looked out the window and there were clear skies and the sun was out.
HAKEEM: Really? Then I guess I was mistaken.

If *It might be that p* is true iff the speaker's knowledge is compatible with *p*, as it is on a simple contextualist theory, then none of this conversation makes any sense. But, argue relativists, it is perfectly natural. A natural move here is to argue that the problem with simple contextualism isn't the contextualism, but the simplicity. Exploring the moves that can be made here will take us too far afield; MacFarlane (2014) goes through a number of possible alternative contextualist theories and shows how examples like the ones involving Kim, Hakeem, and Fabian can be generated to raise problems for each of them. Instead, let's look at how the relativist handles the cases.

According to MacFarlane, relativists are committed to the following principle about retraction:

> The speaker ought to retract the assertion if she has good grounds for thinking that its content is false (as assessed from the perspective she currently occupies).

Given that Kim and Hakeem are evaluating their earlier utterances from new perspectives, perspectives where the assertions are now judged to be false, they both should retract their earlier assertions because they now take them to be false. But arguably this principle overgenerates. Consider this example from von Fintel and Gillies (2008).

ALEX: The keys might be in the drawer.
BILLY: (Looks in the drawer, agitated) They're not. Why did you say that?
ALEX: Look, I didn't say that they were in the drawer. I said they might be there – and they might have been. Sheesh. (von Fintel and Gillies, 2008, p. 81)

The lesson they draw from the example is that retraction is somewhat voluntary. Despite what relativists claim, it is not always true that 'might' claims are retracted or rejected in light of new evidence. Instead, what seems to be true is that solipsistic readings for the modals are virtually always available. They say that the relativist can't easily explain this data. MacFarlane (2014, ch. 10) responds by saying, in effect, that epistemic modal claims sometimes mean what contextualists say they mean, and this is compatible with relativism.

3.2 Disagreement and Agreement

The data about retraction are closely related to a phenomenon that classically was central to debates about relativism. A traditional argument for relativism is that it is necessary to explain 'faultless disagreement.' If one person says Vegemite is tasty, and another says that it is not, the alleged datum is that they are disagreeing, but neither is wrong. Objective treatments of taste save the phenomenon of disagreement, but lose the faultlessness. Contextualist,

or otherwise subjective treatments of taste, save the faultlessness but lose the disagreement. Relativism was alleged to keep both. Faultless disagreement plays a big role in Kölbel (2002), but it has dropped out of the recent literature somewhat. But the focus on disagreement remains, with the central claim being that contextualists cannot explain why some conversations are genuinely disagreements. The following is how MacFarlane puts the point using an example from predicates of personal taste:

> If the truth of my claim that a food is "tasty" depends on how it strikes me, while the truth of your claim that the same food is "not tasty" depends on how it strikes you, then our claims are compatible, and we do not disagree in making them. But it seems that we do disagree – even if we are aware that the source of our disagreement is our differing tastes. (MacFarlane, 2014, p. 8)

This is no argument against any kind of objectivist treatment of predicates of personal taste. The issue is solely whether relativism or contextualism does a better job of explaining the facts about disagreement. We'll look at a couple of reasons for thinking that the problem is less pressing for contextualists than it might first seem, then at an interesting attempt to resolve the problems that remain for contextualism, then at some reasons for doubting that relativism helps solve the remaining problems.

3.3 Clarifying the Data

The puzzle for contextualism is supposed to be that there is a big difference between the felicity of Mark's reply in the following two cases.

SALLY: This chili is tasty.
MARK: I disagree. It's too hot.
SALLY: I'm from Barcelona.
MARK: I disagree. I'm from Oslo.

Assume Sally and Mark are both being sincere. So Sally does like the chili and is from Barcelona, and Mark doesn't like it and is from Oslo. In neither case can Mark sincerely repeat the words that Sally uttered. Indeed, he can sincerely utter the negations of the sentences Sally uttered. But in the first case, this seems to amount to a disagreement, and in the second case it does not. If the context sensitivity of 'tasty' and 'I' is explained the same way, this is mysterious.

But note that there are other examples that do not seem to be amenable to a relativist treatment where Mark can express disagreement with Sally.

SALLY: I like this chili.
MARK: I disagree. It is too hot.

What seems to be going on there is that Mark is disagreeing with an attitude that Sally has, but not with any proposition she expressed. After all, Mark presumably agrees with the proposition that Sally likes the chili. That's why he knows he is disagreeing with her. And that's the content of what he uttered. So some disagreements that are triggered by assertions are not with the content of what is asserted. This might be the germ of an idea for how contextualists can explain the data about disagreement, one nicely developed by Torfinn

T. Huvenes (2012). He argues that in a lot of cases of disagreements involving taste, what we see is not a disagreement about any content, but a disagreement of attitudes.

> Two parties disagree just in case there is something towards which they have conflicting attitudes. This sometimes means that there is a content that one party accepts and the other rejects, but that does not always have to be the case. Just as two parties may have conflicting beliefs, they may also have conflicting desires or preferences. (Huvenes, 2012, p. 178)

If Huvenes is right, then we can't draw any conclusions for semantics from the facts about disagreement.

There is a further problem with using facts about disagreement to argue against contextualism. Consider the following dialogue.

SALLY: Joe might be in Boston.
MARK: I disagree he's definitely in China.

If *Joe might be in Boston* just means *For all I know, Joe is in Boston*, then it is *prima facie* unclear why Mark is disagreeing. After all, it is consistent with Joe's being in China that, for all Sally knows, Joe is in Boston. But, say some contextualists, Mark need not disagree with the whole content of Sally's utterance. He may just be disagreeing with the *prejacent* of the modal, that is, that Joe is in Boston.

Most of the terms of disagreement that have been used in examples motivating anti-contextualism can be seen to target something other than the entire proposition (von Fintel and Gillies, 2008, p. 83). This is a particular problem for arguments for relativism involving epistemic modals. It's possible to say "I disagree," "that's false," "no," or "you're mistaken" and disagree with the prejacent of someone's modal claim, not necessarily the whole claim.

One potential avenue for the relativist to get around this worry is to limit the range of disagreement markers that count as expressing the right kind of disagreement. For example, Tamina Stephenson (2007) limits the terms of disagreement in the examples she uses to "no" and "nuh-uh." And John MacFarlane (2014, p. 11) discusses disagreements that more explicitly target the entire asserted proposition such as "the proposition you expressed is false" and "what you asserted is false." The issue with the two former disagreement markers is that they can explicitly target the embedded clause of an expression. Indeed, Stephenson herself provides a clear example of this phenomenon.

MARY: How's the cake?
SAM: I think it's tasty.
SUE: Nuh-uh, it isn't tasty at all! (Stephenson, 2007, p. 512)

The issue with the disagreement markers that MacFarlane uses is that they have become too technical to do the type of work they need to do. Relativism is meant to be an empirical thesis that relies on natural language data to back it up. We have moved well outside of the realm of natural language and the types of natural language intuitions about the acceptability of sentences that we can get when we move to "the proposition you expressed is false" and "what you asserted is false."

Note that the two contextualist responses we've described here are rather complementary. The point Huvenes makes, that disagreements can involve attitudes other than belief,

seems best served to defuse arguments from disagreements concerning predicates of personal taste. And the point that von Fintel and Gillies make, that disagreements can target prejacents, seems best served to defuse arguments from disagreements concerning epistemic modals. It is possible there are replies to this last point; Weatherson and Egan (2011) for example suggest that examples involving agreements are invulnerable to the response that von Fintel and Gillies make. But as it stands, the dominant trend in the literature seems to be in the direction of thinking the contextualist has the resources to answer these relativist arguments.

3.4 Presuppositions and Common Context

There is a more radical, and perhaps more concessive, response to the disagreement arguments available to the contextualist. Dan López de Sa (2008) develops a contextualist theory that explains the disagreement data by positing that for many context-sensitive terms, there is a presupposition that users of the term are in the same context. (Note that López de Sa calls his theory an "indexical relativist" theory, but it is a kind of contextualist theory in the way the terms are being used here.)

It isn't true in general that there is a presupposition of commonality of context. If two speakers both say "I am happy," they should be interpreted as putting forward different propositions. And that is because, in the relevant sense, they are in different contexts. But, perhaps, many terms are not like this. In particular, in cases where there appears to be a problem with explaining the phenomena involving disagreement, perhaps this is not so. So consider the stock example López de Sa uses, a variant on one that may seem familiar by now.

HANNAH: Homer Simpson is funny.
SARAH: I disagree. Homer is not funny.

If 'funny' denotes a different property when Hannah and Sarah use it, there is no disagreement about propositional content here. And the contextualist account of predicates of personal taste would predict that it could, and perhaps often will, denote a different property on different occasions of usage. But it seems that there is a disagreement here, and arguably even one about propositional content. The solution López de Sa offers is that in any conversation, there is a presupposition that we are applying the same aesthetic standards. The model he uses is a suggestion made by David Lewis (1989) in defending a contextualist treatment of claims about value.

> Wouldn't you hear them saying 'value for me and my mates' or value for the likes of you'? Wouldn't you think they'd stop arguing after one speaker says X is a value and the other says it isn't? – Not necessarily. They might always presuppose, with more or less confidence (well-founded or otherwise), that whatever relativity there is won't matter in this conversation. (Lewis, 1989, p. 84)

Here is how López de Sa develops the point.

> According to the approach, 'is funny' triggers a presupposition of commonality to the effect that both Hannah and Sarah are similar with respect to humor. Thus, in any non-defective conversation where Hannah uttered 'Homer is funny' and Sarah replied 'No, it is not,' it would indeed be common ground that Hannah and Sarah are relevantly alike, and thus that they are

> contradicting each other. After all, provided they are alike, either both Hannah and Sarah are amused by Homer or they are not. (López de Sa, 2008, p. 305)[3]

This is an ingenious idea, but there are a few hurdles to be cleared before it could be declared a full solution to the problem. First, it needs a way to deal with eavesdroppers. If Hannah writes "Homer Simpson is funny" on a scrap of paper, and later Sarah chances upon that paper, she can still say that she disagrees. But it is very odd to think there is a presupposition of commonality of taste with anyone who chances upon one's writings. Second, it needs to account for cases where the presupposition is expressly canceled. The Hannah/Sarah conversation feels natural even in cases where it has been made explicit that Hannah and Sarah have completely different tastes, and they are displaying their differences for the amusement of their friends (MacFarlane, 2014, pp. 131–132). Third, it doesn't quite capture the idea that Hannah and Sarah are contradicting each other. When Sarah disagrees, it shows either that there is a proposition one accepts and the other rejects, or that the context is defective. If we thought the data was that there was a contradiction between what they say, López de Sa's approach can't explain that. This is because López de Sa's approach leaves open the possibility that we have just found ourselves in a defective context. And finally, it isn't obvious how to extend this theory to other terms for which relativism seems promising. It is one thing to say that conversations about humor presuppose a common standard for humor. It is much less plausible to say that conversations about what might be the case presuppose that the parties to the conversation know the same things. None of these hurdles seem impossible to clear, but they do raise doubts about whether presuppositions can solve all the contextualists' problems with disagreement.

3.5 Relativism and Disagreement

If, after all that, we conclude that the contextualist still has a problem with disagreement, it's fair to ask whether the relativist does any better. Let's think again about López de Sa's example of Hannah and Sarah.

HANNAH: Homer Simpson is funny.
SARAH: I disagree. Homer is not funny.

A simple relativist theory assigns truth-conditions to Hannah's utterance relative to a context of utterance, her own, and an index that consists of a world-perspective pair. (Remember that we're using 'perspective' to cover everything other than a world that a relativist may want to put into an index, and that it will be a structured entity.) The content of Hannah's utterance will be set by her context. So it will be (or at least will determine) a function from indices, that is, world-perspective pairs, to truth-values. Call the world Hannah and Sarah occupy w, and their perspectives p_H and p_S. Then the proposition that Homer Simpson is funny will be true relative to $\langle w, p_H \rangle$ and false relative to $\langle w, p_S \rangle$. So there is a proposition Hannah accepts and Sarah rejects, and so they disagree. Doesn't this mean that the relativist can explain the sense in which Hannah and Sarah disagree?

Not so fast. Consider another case involving One, who lives in w_1 where Mars has one moon, and Two, who lives in w_2 where Mars has two moons. They make the following utterances:

ONE: Mars has one moon.
TWO: Mars has two moons.

Now most theorists would say that One and Two have expressed propositions that cannot be true together. (As noted earlier, Schaffer, 2012, disagrees, though not in a way that helps relativism.) But they don't disagree. Among other things, they both think that the other speaks truly.

Hannah doesn't just think that the proposition that Homer Simpson is funny is true relative to $\langle w, p_H \rangle$; she thinks it is simply true. That is because she occupies (for want of a better word) the index $\langle w, p_H \rangle$. And One doesn't just think that the proposition that Mars has one moon is true relative to $\langle w_1, p_1 \rangle$, she thinks it is simply true. But in doing so, she need not be in disagreement with someone who occupies a different index, such as Two, and who thinks it is false. It isn't easy to read off the existence of disagreement from the endorsement of conflicting propositions when the parties occupy different indices. And that raises doubts about whether the relativist has really explained the disagreement. (The Mars example is from MacFarlane, 2014, p. 128. Both Dreier, 2009, and Francén, 2010, raise doubts about whether the relativist can explain disagreement.)

Part of MacFarlane's response to this is to insist that the different elements of an index are treated differently. It is just a fact about disagreement that two speakers who assert incompatible propositions in distinct worlds are not disagreeing, while two speakers who assert incompatible propositions from different perspectives in the same world are disagreeing. And that's related to his view of utterance truth. An utterance is accurate if the proposition expressed by the utterance is true relative to the world it was uttered in, and the perspective it is assessed from. So in Two's context, One's utterance is accurate, even though the proposition One expresses is false in Two's context. On the other hand, in Sarah's context, Hannah's utterance is not accurate. So there is at least a sense in which Hannah and Sarah disagree, although One and Two do not. There is still a lot more work to do to turn this into a full theory of disagreement, and chapter 6 of MacFarlane's book has a very careful study of the varieties of disagreement that are possible on a relativist theory, and how they can be used to explain the data. We're not going to attempt to evaluate the success of these responses, but rather conclude by noting that even if the contextualist has work to do to explain the phenomena involving disagreement, so does the relativist.

3.6 Problems with the Data

Most of the arguments in the literature have started with intuitions about disagreement. We don't think there is anything wrong with this in principle; indeed, it is what we've done so far. But when the intuitions get a little shaky, as they are in a few cases we've described so far, it is worth checking them more carefully, against a broader range of informants. And when that is done, it isn't clear that the data help the relativists as much as the relativists have claimed. Knobe and Yalcin (2014) provide evidence that the following claim, which we'll call (J), isn't as empirically justified as the relativists have made it out to be.

> (J) Competent speaker/hearers tend to judge a present-tense bare epistemic possibility claim (BEP) true only if the prejacent is compatible with their information (whether or not they are the producer of that utterance); otherwise the BEP is judged false.

They argue that many relativists, in particular Egan and MacFarlane, are committed to this claim. Although Egan and MacFarlane differ on several points (Egan takes relativism about propositional truth to be primary, MacFarlane relativism about utterance truth), it does seem true that (J) is important to both of them. As Knobe and Yalcin put it,

> Egan and MacFarlane are both clearly animated by the thought that "people tend to assess epistemic modal claims for truth in light of what they (the assessors) know, even if they realize that they know more than the speaker (or relevant group) did at the time of utterance" (MacFarlane 2011: 160; see also Egan 2007: 2–5, the section entitled "Motivation for relativism: eavesdroppers"). (Knobe and Yalcin, 2014, pp. 3–4)

This isn't what their data showed. The subjects were shown speakers whose evidence strongly, but falsely, suggested that Fat Tony was dead. They overwhelmingly said that an utterance of "Fat Tony is dead" is false, but most said an utterance of "Fat Tony might be dead" was true. (Though it is worth noting that the responses displayed a considerable ambivalence; the answers weren't in line with what either a contextualist or a relativist would straightforwardly predict.) The subjects did say that it would be correct for the speaker who said "Fat Tony might be dead" to retract that utterance once it was clear Fat Tony was alive. But this typically wasn't because they thought the earlier utterance was false.

The point of this study was not to directly target intuitions about disagreement, but rather inter-contextual judgments and felicity of retraction. But the issues are closely related. If subjects who know Fat Tony is alive don't judge that an utterance of "Fat Tony might be dead" is false, then either they don't disagree with such an utterer, or the disagreement is, as Huvenes suggests, not the kind that motivates altering our semantics in the direction relativists suggest.

As Knobe and Yalcin are careful to note, even if relativists were completely wrong about inter-contextual evaluation of utterances, about disagreement, and about retraction, there are still other arguments for relativism. We'll end with one other such argument, due primarily to Tamina Stephenson (2007).

4 Control and Syntax

There is a striking semantic/syntactic phenomenon that epistemic modals and predicates of personal taste share. It's easiest to describe the phenomenon if we assume a contextualist semantics, though we'll eventually use the puzzle to cast doubt on that semantics. So assume, for now, that *It must be that p* is true iff *p* is guaranteed to be true by the (possibly idealized) knowledge of *X*, where *X* is an individual or group supplied by context. (And perhaps the amount of idealization is context-sensitive too.) And assume that *F is tasty* is true iff *F* tastes good to (the possibly idealized version of) *X*, where *X* is again supplied by context.

When 'must' or 'tasty' are not embedded, then *X* seems like it has to include the speaker, and perhaps not much more than that. Even making some other taster or knower salient does not suffice to change the value of *X*. For instance, if one utters "Joe is a great cook and connoisseur of fine food. The meals he prepares are always tasty," the second sentence is not naturally construed as saying that *Joe* always likes the taste of what Joe cooks, but that the speaker and her hearer do (or will). Or if, to borrow an example from Weatherson (2011), we say "Jones must know who the killer is," the relevant *X* for interpreting the 'must' consists of the speaker and her hearers, not Jones. There is something strange about this; it normally isn't that hard to make others relevant in a way that makes them the value of a contextually filled variable.

Or, perhaps, it isn't normally that hard with one key exception. Some context-sensitive terms have, as part of their meaning, that their extension includes the speaker. Such terms

include 'I' and 'we.' Perhaps 'must' and 'tasty' are like 'I' and 'we' in this respect. Except, and here is Stephenson's key insight, there is a big difference between 'I'/'we' and 'must'/'tasty.' The former still pick out the speaker (and perhaps those near her) under the scope of an attitude verb. The latter do not. Consider the natural interpretations of these sentences.

- Joe thinks my gumbo is tasty.
- Joe thinks we must have stolen the candy.

In each case, the personal pronouns ('my,' 'we') have their customary denotations. It is the speaker's gumbo that is being praised, and the speaker and her friends who are being suspected of theft. But 'tasty' and 'must' do not behave like that. It isn't that Joe thinks the speaker likes the taste of her gumbo, but that Joe himself does. And it isn't that Joe thinks the speaker's evidence entails guilt, but that Joe's evidence does.

Stephenson points out that we get even more dramatic results with more complex sentences. Consider these two sentences.

- Mary thinks that Sam thinks it must be raining.
- Mary thinks that Sam must think it is raining. (Stephenson, 2007, p. 490)

The value of *X* for the first 'must' has to be Sam, and for the second 'must' it has to be Mary. In neither case is it the speaker, and in neither case is it even particularly optional how to interpret the 'must.' Compare *Mary thinks that Jane likes her house*, which is naturally read as three-way ambiguous. (The house might be Jane's, or Mary's, or a contextually supplied third party's.) The general point here is that the *X* values that the contextualist posits behave rather unlike other implicit or explicit context-sensitive terms.

The relativist explanation of this is that epistemic modals and predicates of personal taste are, to use Stephenson's phrase, "inherently judge-dependent." That is, they are inherently dependent on some perspective for their truth. In the terms we've been using so far, we need a perspective in the index, and not just in the context, to explain the truth of these claims. When this is combined with a natural view that attitude predicates obligatorily shift the perspective (or judge) parameter, then it just naturally falls out that epistemic modals must take on the perspective of the immediate subjective of the attitude verb. Such a semantics is defended by Stephenson, and by Peter Lasersohn (2005).

It falls out of this semantics that there cannot be "exocentric" uses of bare epistemic modals. These are uses of bare epistemic modals that take on the perspective of someone other than the speaker. (A similar point applies to predicates of personal taste.) This is a nice explanation of the fact that it is very hard to generate these exocentric uses. But it arguably overgenerates; there are cases where it seems we do get the exocentric reading. Here is one case from Egan, Hawthorne, and Weatherson (2005).

> Ann is planning a surprise party for Bill. Unfortunately, Chris has discovered the surprise and told Bill all about it. Now Bill and Chris are having fun watching Ann try to set up the party without being discovered. Currently Ann is walking past Chris's apartment carrying a large supply of party hats. She sees a bus on which Bill frequently rides home, so she jumps into some nearby bushes to avoid being spotted. Bill, watching from Chris's window, is quite amused, but Chris is puzzled and asks Bill why Ann is hiding in the bushes. Bill says, "I might be on that bus." (Egan, Hawthorne, and Weatherson, 2005, p. 140)

The natural reading of what Bill says is that for all *Ann* knows, Bill is on the bus. And this is predicted to be impossible on the type of relativist view that gets us the correct results in the obligatory control cases. Stephenson suggests that there is an ellipsis in Bill's sentence. It should really be understood as:

- Ann is hiding in the bushes because I might be on that bus.

So in a sense, it isn't a bare epistemic modal; it is in a 'because'-clause. Moreover, suggests Stephenson, we should take 'because'-clauses to express something like a person's conscious reasoning or rationale. This means that 'because' acts like an attitude verb and shifts the perspective (or judge) of the epistemic modal. (A similar move, in response to a similar objection, is defended by John MacFarlane, 2014, pp. 272 ff.)

The relativist needs two premises here for the defense to work. The first is that all these exocentric uses either are in the scope of an attitude verb, or are in an explanatory context. The second is that it is fair to treat 'because' as sufficiently like an attitude verb for these purposes. A contextualist could well object to either assumption. But even if they grant both assumptions, a contextualist may want to simply resist the whole line of reasoning. At most what these arguments about control show is that 'might,' and 'tasty,' behave rather differently to other context-sensitive terms. There is nothing inconsistent about just accepting that as a surprising fact. And given how radical a thesis relativism seems to many, accepting relativism as an explanation for these facts about control could well be an excessive reaction. The issues here have not been worked out in nearly as much detail as the issues concerning disagreement and retraction, and we are a long way from having a full accounting of the costs and benefits of the possible dialectical moves.

Notes

1 Strictly speaking, to get MacFarlane's exact view of which kinds of views are relativist we need to complicate things even more. MacFarlane explicitly leaves open the possibility that we have a non-indexical contextualist view about some terms, and a relativist view about others, and expressly says that such a view is a form of relativism. So what we should say is that perspectives are subdivided into those features where the context of utterance is relevant for utterance truth, and those where the context of assessment is relevant to utterance truth. If the first set of features is p_u, and the second set is p_a, then the utterance U of sentence S in context c_u will be true, relative to c_a, iff $f_S(c_u, w(c_u), p_u(c_u), p_a(c_a)) = TRUE$. MacFarlane's own view seems to be that p_u will be empty, so the simplified view we've given in the text is a fair representation of how he thinks actual natural languages work. But there are possible languages where the complications noted here are relevant.

2 The differences between these two formulas will become clear throughout §3.

3 López de Sa (2008) follows Stalnaker (2002) in defining a non-defective conversation as a "[conversation] in which the participants' beliefs about the common ground are all correct. Equivalently, a nondefective [conversation] is one in which all of the parties to the conversation presuppose the same things" (Stalnaker, 2002, pp. 716–717). Following López de Sa, the quote has replaced 'context' for 'conversation' in order to avoid confusion with Lewisian 'contexts.'

References

Cappelen, H., and J. Hawthorne. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cappelen, H., and T. T. Huvenes. 2014. "Relative truth." In *The Oxford Handbook of Truth*, edited by Michael Glanzberg. Oxford: Oxford University Press.

Cappelen, H., and E. Lepore. 2005. *Insensitive Semantics: A Defence of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

DeRose, K. 2009. *The Case for Contextualism: Knowledge, Skepticism and Context*. Oxford: Oxford.

Dreier, J. 2009. "Relativism (and expressivism) and the problem of disagreement." *Philosophical Perspectives*, 23: 79–110.

Egan, A. 2007. "Epistemic modals, relativism and assertion." *Philosophical Studies*, 133(1): 1–22.

Egan, A., J. Hawthorne, and B. Weatherson. 2005. "Epistemic modals in context." In *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, edited by G. Preyer and G. Peter, pp. 131–170. Oxford: Oxford University Press.

Einheuser, I. 2008. "Three forms of truth-relativism." In *Relativising Utterance Truth*, edited by M. Garcia-Carpintero and M. Kölbel, pp. 187–203. Oxford: Oxford University Press.

Evans, G. 1985. "Does tense logic rest on a mistake?" In *Collected Papers*, pp. 343–363. Oxford: Clarendon Press.

von Fintel, K., and A. S. Gillies. 2008. "CIA leaks." *Philosophical Review*, 117(1): 77–98.

Francén, R. 2010. "No deep disagreement for new relativists." *Philosophical Studies*, 151(1): 19–37. DOI:10.1007/s11098-009-9414-6.

Harman, G. 1975. "Moral relativism defended." *The Philosophical Review*, 84(1): 3–22.

Harman, G, and J. Jarvis Thomson. 1996. *Moral Relativism and Moral Objectivity*. Cambridge, MA: Blackwell.

Huvenes, T. T. 2012. "Varieties of disagreement and predicates of taste." *Australasian Journal of Philosophy*, 90(1): 167–181. DOI:10.1080/00048402.2010.550305.

Kamp, H. 1971. "Formal properties of 'now.'" *Theoria*, 37(3): 227–274.

Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Oxford University Press.

Knobe, J., and S. Yalcin. 2014. "Epistemic modals and context: experimental data." *Semantics and Pragmatics*, 7(10): 1–21. DOI:10.3765/sp.7.10.

Kölbel, M. 2002. *Truth Without Objectivity*. London: Routledge.

Kölbel, M. 2004. "Indexical relativism vs. genuine relativism." *International Journal of Philosophical Studies* 12(3): 297–313.

Lasersohn, P. 2005. "Context dependence, disagreement and predicates of personal taste." *Linguistics and Philosophy*, 28(6): 643–686.

Lewis, D. 1979. "Attitudes *de dicto* and *de se*." *Philosophical Review*, 88(4): 513–543.

Lewis, D. 1980. "Index, context, and content." In *Philosophy and Grammar*, edited by S. Kanger and S. Öhman, pp. 79–100. Dordrecht, Netherlands: Reidel.

Lewis, D. 1989. "Dispositional theories of value." *Proceedings of the Aristotelian Society*, suppl. vol. 63: 113–137. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 68–94. Cambridge: Cambridge University Press, 2000.

López de Sa, D. 2008. "Presuppositions of commonality: an indexical relativist account of disagreement." In *Relative Truth*, edited by M. Garcia-Carpintero and M. Kölbel, pp. 297–310. Oxford: Oxford University Press.

MacFarlane, J. 2014. *Assessment Sensitivity*. Oxford: Oxford University Press.

Recanati, F. 2007. *Perspectival Thought: A Plea for (Moderate) Relativism*. Oxford: Oxford University Press.

Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Schaffer, J. 2012. "Necessitarian propositions." *Synthese*, 189(1): 119–162. DOI:10.1007/s11229-012-0097-8.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, R. 2002. "Common ground." *Linguistics and Philosophy*, 25(5–6): 701–721.

Stephenson, T. 2007. "Judge dependence, epistemic modals, and predicates of personal taste." *Linguistics and Philosophy*, 30(4): 487–525.

Weatherson, B. 2011. "No royal road to relativism." *Analysis*, 71(1): 133–143. DOI:10.1093/analys/anq060.

Weatherson, B., and A. Egan. 2011. "Epistemic modals and epistemic modality." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 1–18. Oxford: Oxford University Press.

PART II

Reference, Identity, and Necessity

31

Modality

BOB HALE

1 Preliminary Considerations: Philosophical Issues

1.1 The Importance of Modal Notions

The notions of necessity and possibility, of what must be so and what may be so, and the derivative notion of contingency – of what is so but might be otherwise – are ones which very few philosophers find themselves able to do without. It is, to take one arguably fundamental case, hard to see how an adequate explanation of the notion of valid argument, as distinct from that of proof in a specified formal system, might run, save in terms of the idea that the conclusion *must* be true if the premises are. Even those vigorously skeptical of modal notions seem unable to voice their skepticism without recourse to them. When Quine denies that there are any statements immune from empirical revision – any necessarily true statements, as he construes the notion – he is not claiming that any statement accepted at any time is one which we *will* at some time in fact reject; what he is denying is the existence of statements which we *could* not be led to reject. It is difficult to see how his skepticism about necessity could be so much as expressed without employing the notion of possibility.[1] And once a notion of possibility has been granted house-room, the intelligibility of a correlative notion of necessity can hardly be denied. It thus appears that philosophical skepticism about necessity must, if it is not to fall into incoherence, take the form of denying the existence of truths having that character, rather than rejecting the notion altogether. That is not, of course, to deny that the notions of necessity and possibility stand in much need of elucidation; on the contrary, it is surely a central task of a philosophy of modality to provide an account of them.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

1.2 *Relative and Absolute Modalities*

As a first step, we may usefully begin by drawing some distinctions among different notions of necessity and possibility. Probably the single most important distinction to be drawn is between *absolute* and *relative* kinds or senses. Roughly speaking, the distinction here is between a truth's being necessary outright or without qualification, and its being a necessary consequence of some pre-assigned collection of statements which are taken to be true, but are not (either necessarily or even typically) themselves true by necessity. When philosophers speak of broadly logical necessity or (what is not to be assumed the same thing) metaphysical necessity, they intend an absolute sense of necessity; but when they speak of natural necessity, physical necessity, biological necessity and the like, it often appears to be relative necessity that they have in mind. What is usually meant by saying that it is, say, physically necessary that p is that it follows from the laws of physics that p; saying that it is physically possible that p is saying that it is consistent with the laws of physics that p. Since the laws of physics – which need not, of course, be as we suppose them to be – are certain true propositions belonging to physics, a proposition cannot be physically necessary without being true; but unless the laws of physics are themselves held to be absolutely necessary, what is physically necessary will not normally be necessarily true in any absolute sense. Similar remarks apply to other relative notions, of course. Clearly, whenever we have a more or less definite body of propositions constituting a discipline D, there can be introduced a relative notion of necessity – expressible by 'It is D-ly necessary that' – according to which a proposition will be D-ly necessary just in case it is true and a consequence of D.[2]

The modal verbs 'must' and 'may' are also commonly employed in epistemic senses, as when we say things like 'He must have got off the train at Oxenholme' (when we know that he was aboard when the train left Penrith and was not on it when it arrived at Lancaster) or 'The train may have been delayed.' Such uses may sometimes be correctly explained as involving relative notions – it being epistemically possible that p if it is consistent with what we know that p, and epistemically necessary that p if it follows from things we know that p.[3] But it seems clear that this is not right for all cases: when we assert that, for all we know, Goldbach's Conjecture may be true, but may equally be false, we are not claiming that neither the Conjecture nor its negation is deducible from number-theoretic propositions which we take ourselves to know, but making the far more modest claim that thus far no one has succeeded either in proving that every even number is the sum of two primes or in finding a counter-example.

The notion of consequence employed in characterizing any relative notion of necessity is plausibly taken to be that of broadly logical consequence.

Since, if there are any statements at all that deserve to be regarded as logically necessary, statements recording the connection between the premises and conclusion of a valid inference are surely among them,[4] our characterization of relative necessity assumes that some truths are broadly logically necessary. It is, however, hard to see how logical necessity itself can be other than an absolute notion. For logical truths to be merely relatively necessary, there would have to be some further truths of which they are (logical) consequences. But first: What could these truths be? Aren't logical truths precisely those which are consequences of the null set of premises? And second: Supposing there is a set K of truths of some other sort, of which logical truths are consequences, what are we to say of the conditionals 'If K then p' where p is any logical truth? How can these conditionals be other than logical truths which are absolutely necessary? It thus appears that, from a

philosophical standpoint, it is the absolute notions that are fundamental. The various relative notions presuppose and are to be explained in terms of them, and it is on them that we shall concentrate here. Accordingly, all subsequent references to necessity and possibility are to absolute notions unless otherwise indicated. In addition, discussion will be largely concentrated upon necessity, although much of what is said will apply, with more or less obvious adjustments, to possibility.

1.3 Comparability and Strength of Absolute Modalities

The foregoing remarks leave it open that there may be more than one absolute notion of necessity. If there is more than one – as would be the case, for example, if logical and metaphysical necessity are distinct notions and both are absolute – this gives rise to questions of comparability and relative strength. In saying that two notions of necessity – necessity$_1$ and necessity$_2$ – are comparable, I mean simply that either 'it is necessary$_1$ that *A*' entails 'it is necessary$_2$ that *A*' whatever statement *A* may be, or conversely. It is clearly a substantial claim that two notions of necessity are comparable. It is more often simply assumed rather than argued that logical necessity is the strongest kind. If attention is restricted to absolute notions and their comparability is assumed, it is hard to see how the common view could seriously be questioned. For if there is an absolute notion of necessity, necessity*, which is distinct from but comparable with logical necessity, either 'It is necessary* that *p*' entails 'It is logically necessary that *p*' or conversely. And likewise for the corresponding notions of possibility: either 'It is possible* that *p*' entails 'It is logically possible that *p*' or conversely. If necessity* is stronger than logical necessity, then possibility* is weaker than logical possibility and so does not entail it. But since it is logically possible that *p* provided that *p* entails no contradiction, this can only be so if it is possible* for a contradiction to be true. The current vogue for 'true contradictions' notwithstanding, it seems to me that there is no good sense of 'possible' for which this holds. If comparability is not assumed, however, the claim that logical necessity is the strongest kind of necessity calls for further argument.

An argument to the purpose was put forward by Ian McFetridge in a posthumously published essay on logical necessity (McFetridge, 1990). As his original argument involves some avoidable complications, I shall give my own version. If there is a kind of absolute necessity than which logical necessity is not stronger, there will likewise be a corresponding kind of possibility than which logical possibility is not weaker; and for this notion of possibility, it will hold that it is possible for the premise(s) of a valid argument to be true but its conclusion false.[5] Let us use '*Poss A*' to express that it is, in this sense, possible that *A*. We make the following five assumptions about entailment and '*Poss*' (I write $A \Rightarrow B$ for '*A* entails *B*,' understood as meaning that there is a valid argument from *A* to *B*):

(A1) If $A \Rightarrow$ B then $A \wedge C \Rightarrow B$ *and* $C \wedge A \Rightarrow B$
(A2) $A \Rightarrow A$
(A3) If $A \Rightarrow B$ and $A \Rightarrow C$ then $A \Rightarrow (B \wedge C)$
(A4) If *Poss A* and $A \Rightarrow B$ then *Poss* B
(A5) $\neg$*Poss* $(A \wedge \neg A)$

On these assumptions, we can show that if $A \Rightarrow B$ then $\neg$*Poss* $(A \wedge \neg B)$:

1	(1) $A \Rightarrow B$	Assumption
2	(2) *Poss* $(A \wedge \neg B)$	Assumption

1	(3) $(A \wedge \neg B) \Rightarrow B$	(1) by (A1)
	(4) $\neg B \Rightarrow \neg B$	by (A2)
	(5) $(A \wedge \neg B) \Rightarrow \neg B$	(4) by (A1)
1	(6) $(A \wedge \neg B) \Rightarrow (B \wedge \neg B)$	(3), (5) by (A3)
1,2	(7) *Poss* $(B \wedge \neg B)$	(2), (6) by (A4)
	(8) $\neg$*Poss* $(B \wedge \neg B)$	by (A5)
1	(9) $\neg$*Poss* $(A \wedge \neg B)$	(2), (7), (8) by *reductio ad absurdum*

Remember that *Poss* represents *any* sense of 'possibly' conforming to (A4) and (A5), so that if our ancillary assumptions about entailment are correct, it follows that there is no such (absolute) sense of 'possibly' in which it is possible that the premise(s) of a valid argument should be true but its conclusion false. I do in fact hold all five assumptions to be met by entailment and any reasonable notion of possibility. Obviously anyone wishing to impose a quite strong relevance-constraint upon entailment will find (A1) unacceptable; and paraconsistent logicians may feel free to reject (A5). It seems, nevertheless, to be a result of some interest and importance that, without assuming comparability, it can be shown from assumptions which, whilst not quite indisputable, are not grossly immodest, that logical necessity is the strongest absolute notion of necessity.[6]

1.4 The Philosophical Problem of Necessity

What should a philosophical account of the (absolute) notion(s) of necessity accomplish? Michael Dummett provided what many ought to find an apt formulation of the task confronting us:

> The philosophical problem of necessity is twofold: what is its source, and how do we recognize it?[7]

As pinpointing what has been the preoccupation of much philosophical discussion of necessity, Dummett's formulation can scarcely be faulted. Plainly, however, Dummett's questions carry presuppositions which both can be, and actually have been, called into question. One is that there is indeed such a thing as necessity – that the notion of necessity has application. Another, which may still be questioned by one who grants the first presupposition, is that the notion has application in such a way as to give rise to a genuine epistemological problem: in effect, that there is a distinctive class of *truths* about what is necessary – truths of the form 'It is necessarily true that *p*' – concerning which it is to be enquired how we (can) know them. These presuppositions will occupy us for the remainder of this chapter.

2 Quine's Skepticism and Reactions to It

2.1 Quine's Solution to the Problem of Necessity

The burden of Quine's celebrated attack upon the 'first dogma' of empiricism is that there are no statements immune to revision in the face of recalcitrant experience, and so no statements which are analytic in the sense of holding true, come what may. Since Quine makes no distinction between the claim that a statement is analytic and the claim that it is necessary – holding, as he does, that the problem of making sense of the adverb 'necessarily' is

one and the same with that of achieving a satisfactory explanation of 'analytic' – his attack on the notion of analyticity is simultaneously one upon that of necessity, and his eventual denial that there are any statements which are true come what may is thus a denial that there are any necessary truths. He may thus be seen as rejecting both presuppositions of Dummett's formulation of the problem of necessity. Equally well, he may be seen as offering a negative solution to it. Earlier and contemporary empiricists, including the logical positivists (cf. Ayer, 1946, ch. 4), had accepted that there are necessary truths, knowable *a priori*, and thus confronted Dummett's problem in spades: how to explain, compatibly with their central thesis that all genuine knowledge derives from sense-experience, first, how there could be such truths and, second, how they could be known. Quine's solution was as dramatic as it was radical: there are no necessary truths, hence no satisfiable demand for an account of their source, nor for an explanation how they are known.

2.2 Quine's Skepticism Finessed?

It may well seem that if Quine is right then necessity is, to borrow some words from Hilary Putnam, yet another example of a subject without an object. And that it surely would be, if Quine were right in thinking that there are no statements that are true come what may; for what are necessary truths, if not that? Interestingly, Putnam himself argues, in the paper from which the words are borrowed,[8] that this conclusion is too swiftly drawn. We can grant that Quine's attack is entirely destructive of a certain 'epistemic' notion of necessity – the conception of a true statement as necessary iff not liable to rational revision – and yet still find application for another philosophically important, but non-epistemic, notion. What he has in mind is what many philosophers, following Kripke (and Putnam himself), have called metaphysical necessity.[9] The presently relevant point of this shift in thinking about necessity, as Putnam seems here to conceive it, is that it detaches the notion of necessity from that of being knowable *a priori* (which Putnam, like Quine, understands as requiring unrevisability). According to this way of thinking, it is metaphysically necessary, for example, that water is H_2O (that is, there is no possible world in which water exists but is not this compound of hydrogen and oxygen);[10] but this necessity is something we know – and can only know – *a posteriori*. And because the claim of necessity does not entrain apriority, it escapes Quine's attack.

Whatever may be said in favor of the notion of metaphysical necessity (or in favor of making space for the conception of necessities knowable only *a posteriori*), there is an obvious difficulty with this way of rescuing the topic of necessity from Quine's clutches. This emerges as soon as we reflect on how, in more detail, *a posteriori* necessities are supposed to be known. At least if we follow Kripke's own suggestion – and no one, to my knowledge, has proposed anything better, or significantly different[11] – our *a posteriori* knowledge that water is necessarily H_2O is the output of a *modus ponens*, applied to two other pieces of knowledge as premises: that water is H_2O, and that if water has a certain chemical composition, it has that chemical composition of necessity (that is, its chemical nature is essential to it). On anyone's view, our knowledge of the minor premise cannot but be *a posteriori*, and this is why our knowledge of the conclusion is so too. But in Kripke's view, the conditional major premise is known by 'philosophical analysis' (Kripke, 1971, p. 153) – and is thus, presumably, a necessary truth, known *a priori*. If that is right, the possibility of *a posteriori* necessities rests squarely on the shoulders of *a priori* necessities; it therefore appears to be a complete illusion that Quine's attack on the latter has been finessed.[12]

2.3 More Direct Defenses against Quine's Challenge

It thus appears that if we are to take necessity seriously, we have no option but to confront Quine's challenge directly. Full discussion of this issue is beyond the scope of this chapter; but two possible lines of counter-attack may be briefly reviewed.[13] First, it may be argued that Quine's own position – an uncompromisingly global empiricism in which all statements we accept have the status of empirical hypotheses, up for revision or retention in the light of experience, with the choice being guided by broadly pragmatic considerations[14] – is itself untenable as a direct consequence of his refusal to accord *a priori* status to any statements whatever. An argument to this general effect – which, so far as I know, has yet to receive an effective counter – has been advanced by Crispin Wright (q.v. Wright, 1986). Wright's central claim is that Quine's position is ultimately unstable because viciously regressive. Among the statements which, in Quine's view, must be up for revision in any (dis)confirmation situation will be certain conditionals purporting to record the logical consequences of our currently accepted combination of empirical theory plus underlying logic. In other words, one option when we are confronted with some recalcitrant sequence of experiences will be to retain our threatened empirical theory along with its underlying logic, eliminating recalcitrance by way of rejecting the claim that their combination does, after all, have the troublesome consequences.[15] When, if ever, should we exercise this particular option? Well, presumably, when doing so results in the optimal balance between minimizing clashes with experience and maximizing simplicity and economy of overall theory. To arrive at a rational assessment on that matter, we must see how that option fares in comparison with the various others available. This will involve, *inter alia*, judgments about what the observational consequences are of the various options. But these in turn will require deploying some hypotheses about the logical consequences attending implementation of the different options. With which hypotheses about their logical consequences should we work? Well, clearly we should work with the *best* such hypotheses. And we should decide what those are by applying Quine's pragmatic criteria; but it was precisely in the course of trying to apply those criteria that we were led to our present question. If we have not come full circle, then we have set off on a vicious infinite regress. If we are ever to be able to apply Quine's pragmatic guidelines, some statements must be kept out of the pragmatic melting pot and treated as not being up for empirical revision. If not, then, as Wright in one place puts it, "the pragmatic methodology is drained of all directive content" (Wright, 1986, p. 222).

Second, and independently of the preceding counter, a defender of necessity might preserve the linkage between it and apriority[16] by severing that (implicit in Quine and explicit in Putnam) between apriority and (absolute) unrevisability. There is no evident reason why we should take ourselves to be immune from error in the *a priori* detection of necessary truths; on the face of it, having *a priori* grounds for believing that *p* is one thing, and being infallible about the matter is another. (See Chapter 23, ANALYTICITY, §I for some further discussion of apriority.)

3 Modal Realism 1: Realism about Possible Worlds

As already remarked, it is a further presupposition of Dummett's formulation of the problem of necessity that necessary truths constitute a distinctive class of truths: that there are, in some sense, genuinely modal facts, not reducible to facts of any other kind. In this sense,

Dummett's formulation presupposes a realistic attitude towards modality. And as also remarked, realism in this sense has not gone unchallenged. Opposition to it goes at least as far back as Hume, who denied that necessity is anything to be detected among the objects of knowledge, maintaining instead that it is nothing but the projection of our own sentiments or attitudes, themselves induced by patterns in our experience (see Hume, 1960, Bk.1, Sec. xiv). Hume, of course, had causal or natural necessity in view. But some, following Hume's lead quite closely, have sought to extend his approach to modality in general (cf. Blackburn, 1984, pp. 210–217; 1986; Craig, 1985). Others have evinced less sympathy with Hume's projectivist explanation, but have endorsed the non-cognitivism about necessity on which it builds.[17] We shall return to this line of anti-realist theorizing. I want first to discuss a different form the realism it opposes may assume which entails, but goes appreciably beyond, the comparatively modest variety just sketched, and which has been the focus of much recent discussion: realism about possible worlds.

3.1 Modal Logic and Possible-World Semantics

Before we proceed with our discussion of this type of realism, a brief outline of the main ideas of modal logic and the so-called possible-world semantics developed for it in the 1950s and 1960s may be useful.

3.1.1 Modal Logics

A language for modal logic is obtained by adding some modal operators – usually, two unary sentential operators, $\Box$ and $\Diamond$, read as 'necessarily' and 'possibly' – to the underlying language for non-modal logic (which is usually first-order, but may be higher-order). A system of modal logic is then obtained by adding axioms, or rules of inference, governing these new operators. Depending upon what axioms or rules are chosen, weaker or stronger modal logics result. The weakest system, known as K, is obtained from the underlying non-modal logic by adding $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ as an axiom and a rule of necessitation which allows us to assert $\Box A$ as a theorem whenever A is a theorem. If we understand the necessitated, or strict, conditional as asserting that A entails B, then the axiom says that what is entailed by a necessary truth is itself necessarily true. A stronger, but still quite weak, system, known as T, is obtained by adding the further axiom $\Box A \rightarrow A$, which says that what is necessarily true is true. The operator $\Diamond$ is usually taken to be interdefinable with $\Box$. If $\Box$ is taken as primitive, $\Diamond$ may be defined $\Diamond A = def \neg\Box\neg A$. If $\Diamond$ is primitive, $\Box$ can be defined $\Box A = def \neg\Diamond\neg A$; in this case, the axioms will be adopted in their equivalent forms governing possibility (for example, the T axiom becomes $A \rightarrow \Diamond A$). Stronger systems are obtained by adding further axioms, which may seem less intuitively obvious. Two of the most important are $\Box A \rightarrow \Box\Box A$ and $\Diamond A \rightarrow \Box\Diamond A$.[18]

The first of these, which asserts that what is necessary is necessarily necessary, is the characteristic axiom of the system S4 – the system which results from adding this axiom to those for the system T. The second, which asserts that what is possible is necessarily possible, is the characteristic axiom of the yet stronger system S5 – which results from adding this axiom to those for T. A third is $A \rightarrow \Box\Diamond A$, known as the Brouwersche axiom, which, if added to the axioms for the system T, gives another system, B, which, like S4, is intermediate in strength between T and S5, but disjoint from, and neither stronger nor weaker than, S4.

The characteristic axioms for S4 and S5 involve what are known as *iterated* modalities – that is, unbroken sequences of modal operators such as occur in the formulae $\Box\Box A$ and $\Box\Diamond A$.

Since the converses of these axioms – *viz.* $\Box\Box A \rightarrow \Box A$ and $\Box\Diamond A \rightarrow \Diamond A$ – are already theorems of the weaker system T, the effect of the axioms is to ensure the equivalence of the iterated modalities involved to simple modalities. Thus in S4, any finite sequence of $\Box$s can be simplified to a single $\Box$, and in S5 any sequence $\Box\Diamond$ simplifies to $\Diamond$. Thus adding these axioms reduces the number of inequivalent modalities. Since the S4 axiom is a theorem of S5, both kinds of simplification can be effected in the latter system. For example, $\Box\Box\Diamond A$ simplifies to $\Diamond A$; and since $\Diamond\Diamond A \rightarrow \Diamond A$ and $\Diamond\Box A \rightarrow \Box A$ are theorems of S4 and S5 respectively, $\Diamond\Diamond\Box A$ simplifies in S5 to $\Box A$. Maximum simplicity is achieved in S5. By a *modality* is meant an unbroken sequence of zero or more of the modal operators and negation. In S4, there are just 14 inequivalent modalities, *viz.* _ (the null modality), $\Box$, $\Diamond$, $\Box\Diamond$, $\Diamond\Box$, $\Box\Diamond\Box$, $\Diamond\Box\Diamond$, together with their negations $\neg$, $\neg\Box$, and so on. In S5, these reduce to six, *viz.* _, $\Box$, $\Diamond$, and their negations. In T, by contrast, no simplification is possible, so that there are infinitely many distinct modalities.[19]

If the axioms for a given modal system are added to a system of first- or higher-order quantification theory, we obtain a system of quantified modal logic – quantified K, or quantified S5, say. The interaction of modal operators with quantifiers produces interesting results and gives rise to disputed questions about what principles involving such interactions should hold. Two particularly interesting and controversial such principles are the Barcan Principle $\forall x\Box A \rightarrow \Box\forall x A$ and its converse, $\Box\forall x A \rightarrow \forall x\Box A$. Which, if any, of these principles is a theorem depends on both the underlying non-modal logic and the modal axioms adjoined to it. If the underlying logic is classical, Converse Barcan is provable in quantified K, and so in any stronger system; and Barcan is provable in quantified B, and so in quantified S5, which includes B. Barcan is controversial, at least in part, because it seems, at least to some thinkers, that it could be true that everything which actually exists is necessarily thus-and-so, but also true that there might have existed something which is not thus-and-so. One reason why Converse Barcan is controversial can be seen by considering its instance: $\Box\forall x\exists y\ x = y \rightarrow \forall x\Box\exists y\ x = y$. The antecedent here is itself a theorem of quantified K, since it follows from the Reflexivity of Identity $\forall x\ x = x$ by existential generalization and the Rule of Necessitation. So the consequent is likewise a theorem. But the consequent is naturally understood as asserting that everything exists of necessity, contrary to the widespread belief that there are many things which might not have existed. Neither Barcan nor its Converse are provable, even in quantified S5, if the underlying logic is free (see Glossary).

Questions about which modal principles – the K, T, S4, B, and S5 axioms, for example, and the Barcan and Converse Barcan principles – should be among the theorems of a modal logic have attracted a good deal of discussion. Such questions have little or no point unless asked with respect to some proposed interpretation of the modal operators. If $\Box A$ is interpreted *alethically*, so that it expresses that it is *necessarily true* that *A*, then it is clear that at least the K and T axioms should hold, but there is more scope for argument about the stronger B, S4, and S5 principles. If, instead, $\Box$ is interpreted *deontically*, as expressing that it *ought to be the case* that *A*, or *doxastically*, as expressing, say, that it is *reasonable to believe* that *A*, then it should not conform to the T axiom, even if it should satisfy the K axiom. And of course, other interpretations are possible. Just as 'may' and 'must' are often used to express epistemological claims (e.g., 'She may have missed the train,' 'I must have left my wallet in my other jacket pocket'), so $\Box$ and $\Diamond$ may be interpreted epistemically, to express claims about what may or must be the case, relative to a certain body of knowledge. Further, different alethic interpretations are possible, so that $\Box$ may be taken to express logical necessity, or metaphysical necessity, or natural necessity, to mention only the most obvious alternatives.[20]

3.1.2 *Possible World Semantics*

It comes very easily to us to express modal thoughts in terms of possible worlds – what is necessarily true is what holds true not only in the actual world, but in all other possible worlds as well; what is possibly true is what holds true in some possible world – perhaps the actual world, but perhaps some merely possible world.[21]

These easy paraphrases form the starting point of possible-world semantics for modal logics, first developed in the late 1950s and early 1960s by Saul Kripke and others. Broadly speaking, an interpretation of an otherwise standard first-order language to which modal operators are added consists of a set, or domain, W, of possible worlds (with one of them, w_α designated as the actual world) and a set, or domain, I, of individuals, together with functions which assign subsets I_w of I to the elements of W, elements of I to the individual constants (if any) of the language, and, relative to each element of W, subsets of I_w^n to the n-place predicates of the language. Under more or less obvious stipulations for dealing with connectives, quantifiers, and modal operators, we can then define what is required for a formula of the language to be true at a world w in W. A formula is true in the interpretation if true at w_α and valid if true in all interpretations. In this setting, the condition for $\Box A$ to be true at w_α is that A be true at each w in W, while that for $\Diamond A$ to be true is that A be true at some w in W. Thus $\Box$ and $\Diamond$ are, in effect, interpreted as universal and existential quantifiers over worlds.

A semantical account along these lines *may* be viewed as no more than an algebraic or model-theoretic device, in relation to which metalogical results about the soundness, completeness and so forth of a specified system of modal logic may be established (see Chapter 35, REFERENCE AND NECESSITY, §2). From such a standpoint we are under no stronger pressure to take the informal accompanying patter about possible worlds as aspiring to literal truth than we are so to regard talk of truth-values intermediate between truth and falsehood as it occurs in the many-valued interpretations by means of which, say, independence results are established. If, however, we are disposed to view possible-world semantics as real semantics – that is, as furnishing genuinely illuminating statements of truth-conditions for modal sentences, and perhaps as forming the basis of explanatory accounts of other concepts, then, or so it seems, we must take the possible-world semantics as attempting, in Alvin Plantinga's memorable phrase, "to spell out the sober metaphysical truth about modality" (Plantinga, 1974, p. 125): we must take seriously the idea that modal statements are, in effect, to be construed as quantifications over a domain of real entities comprising possible worlds. No one has defended such a realistic attitude towards possible worlds with greater ingenuity and resourcefulness than David Lewis. A full-scale discussion of his position lies well beyond the scope of the present chapter, which must settle for a brief and selective review of the main arguments advanced on either side, and of some of the alternatives to Lewis's uncompromising realism which have been canvassed.

3.2 *Arguments for Realism about Worlds*

3.2.1 *The Paraphrase Argument*

In an early defense of realism, Lewis gives some prominence to an argument which represents the central thesis of his realism – that there are possible worlds other than the one we happen to inhabit – as no more than an innocent paraphrase of a very general modal belief from which he expects no one to dissent:

> It is uncontroversially true that things might have been otherwise than they are…. Ordinary language permits the paraphrase: there are many ways things could have been besides the way

> they actually are. On the face of it, this sentence is an existential quantification. It says that there are many entities of a certain description, to wit 'ways things could have been.' I believe that things could have been different in countless ways; I believe permissible paraphrases of what I believe; taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called 'ways things could have been.' I prefer to call them 'possible worlds.' (Lewis, 1973, p. 84)

Given Lewis's insistence, equally prominent in this early discussion, and repeated in subsequent defenses, that other possible worlds are (just) more things of the same kind as the actual world, it would seem to follow that the actual world is one of the ways things could have been – the way they are. This strongly suggests what might be called a Tractarian conception[22] of possible worlds, as collections, probably maximal, of possible states of affairs, the actual world being thought of as that collection of possible states of affairs, each of which is realized.

Whatever merit this argument might be thought to possess, it is clear that it can provide absolutely no support for the version of realism which predominates even in Lewis (1973), and on which Lewis stabilizes in subsequent writings (cf. especially Lewis, 1986, ch. 1), according to which worlds are spatio-temporally and (therefore) causally closed systems, typically largely populated by concrete entities of various kinds, the actual world being identified not as a certain collection of states of affairs, but with one particular such system of concrete (and possibly also abstract) entities, comprising – as Lewis charmingly puts it – Lewis and all his surroundings.[23] It is to this rival conception that Lewis appeals (only paragraphs later than the one in which the quoted argument appears), in objecting to the view that (merely) possible worlds are maximally consistent sets of sentences:

> given that the actual world does not differ in kind from the rest, [this view] would lead to the conclusion that our actual world is a set of sentences. Since I cannot believe that I and all my surroundings are a set of sentences … I cannot believe that other worlds are sets of sentences either.[24]

Lewis is chary of saying outright whether worlds themselves are concrete entities, pending clarification of the abstract/concrete distinction; but he is confident that, if the distinction can be satisfactorily elucidated at all, it will not be found that the actual world falls on one side of the divide and other possible worlds on the other.

3.2.2 Arguments from Explanatory Virtue

It is hardly surprising that the Paraphrase Argument disappears from view in Lewis's later defenses of his realism, as does the suggestion that realism involves no serious departure from 'common opinion' (cf. Lewis, 1986, p. 133, also p. 100). Argument of a quite different stripe comes to center stage – argument broadly to the effect that realism about worlds should command our acceptance by dint of its distinctive explanatory advantages. These, in Lewis's view, are many, various, and substantial. Fundamental, of course, must be the claim that by understanding ordinary modal claims as, in effect, quantifications over a domain of possible worlds, we gain an illuminating account of their truth-conditions. Building on this, it may be claimed that a satisfying explanation can be given of uncontroversial facts about validity and invalidity of modal inferences. A simple illustration is afforded by the patently invalid inference from $\Diamond p$ and $\Diamond q$ to $\Diamond(p \wedge q)$. When modal operators are construed as, in effect, quantifiers over a domain of possible worlds, this inference assumes the form: $\exists w$

True(p, w), $\exists w$ *True*(q, w) $\therefore$ $\exists w$ *True*($p \wedge q$, w), the invalidity of which is then readily recognizable as a special case of the generally invalid quantificational pattern: $\exists xA$, $\exists xB \therefore \exists x(A \wedge B)$. Likewise, the validity of $\Box p$, $\Box q \therefore \Box(p \wedge q)$ reappears as a special case of the quantificational validity $\forall xA$, $\forall xB \therefore \forall x(A \wedge B)$. Lewis is at pains to stress what he sees as the advantages of his realism as a source of good explanations of other philosophically important and perhaps otherwise problematic concepts, besides those involved in modal logic, narrowly conceived. A prominent example is the subjunctive or counterfactual conditional 'If it were the case that p, it would be the case that q,' whose truth-condition on Lewis's account is, roughly, that q holds true at all those possible worlds at which p holds which are 'closest' to (i.e., most like) the actual world. Another is the analysis of propositions as sets of possible worlds, under which the proposition that p is identified with the set of possible worlds at which it is true that p. In general terms, the thought is that these and other explanatory advantages cannot be enjoyed without embracing realism about possible worlds themselves.[25]

3.3 Alternatives to and Arguments against Realism about Worlds

There are two fairly obvious ways in which this kind of case for realism may be countered. First, it may be granted that possible-worlds apparatus does indeed bring distinctive explanatory advantages, but argued that these advantages can be enjoyed without engaging in the full-blooded realism about possible worlds which Lewis advocates. Second, it may be argued that the apparent explanatory advantages of possible-world semantics are illusory, because alternative explanations can be provided which make no essential play with possible worlds at all. We shall briefly review some of the main lines of thought which have been advanced under these two broad headings.

Under the first, one early reaction to Lewis's uncompromising or 'extreme' brand of realism was Robert Stalnaker's moderate realism. Stalnaker takes Lewis's realism to consist of four theses: in his own words, they are:

(1) Possible worlds exist
(2) Other possible worlds are things of the same sort as the actual world – "I and all my surroundings"
(3) The indexical analysis of the adjective 'actual' is the correct analysis[26]
(4) Possible worlds cannot be reduced to something more basic (Stalnaker, 2003, p. 27)

It is, Stalnaker claims, thesis (2) which "gives realism about possible worlds its metaphysical bite, since it implies that possible worlds are not shadowy ways things could be, but concrete particulars, or at least entities which are made up of concrete particulars and events" – so that "even a philosopher who had no qualms about abstract objects like numbers, properties, states and kinds might balk at this proliferation of full-blooded universes which seem less real to us only because we have never been there" (Stalnaker, 1976, p. 68). His moderate realism results from rejecting thesis (2), whilst retaining the other three. Rejecting thesis (2) allows us to preserve the identification of the actual world with David Lewis and his surroundings, whilst viewing other possible worlds as no more than ways things might have been, and thus as things of a quite different kind from the actual world. Stalnaker is less than fully explicit on what precisely 'ways things might have been' are, but his view appears to be that they are properties (and so not collections of systems of concrete objects, such as the actual world is).[27] Since properties can be held to exist uninstantiated, this leaves us free

to maintain that (merely) possible worlds exist; "that there really are many ways that things could have been – while denying that there exists anything else that is like the actual world" (Stalnaker, 1976, p. 68).

Whatever ontological advantage may be thought to attach to maintaining that merely possible worlds are entities of a radically different kind from the actual world, it appears to have some awkward consequences, at least for anyone who wishes to regard possible-world semantics as providing an illuminating account of the truth-conditions of modal statements. An immediate difficulty arises over the interpretation of modal operators as quantifiers over possible worlds. Since merely possible worlds are to be thought of as properties, it follows that they cannot, as is usually supposed, be first-order quantifiers over objects, but must be (at least) second-order quantifiers over properties.[28] This might, by itself, be thought to render moderate realism a significantly less attractive option (and would, of course, be seen as a fatal flaw by those who view higher-order quantification with no less suspicion than they do Lewisian worlds). But there is a more straightforward difficulty which is quite independent of any hostility towards higher-order quantification. If world quantifiers range over a homogeneous domain of properties, that domain cannot include the actual world; with the result that the semantics will fail to validate such obviously valid inferences as those from a proposition's necessary truth to its truth *simpliciter* (that is, truth in the actual world), and from its truth to its possible truth. This problem was first noticed, so far as I know, by Colin McGinn,[29] who suggests that Stalnaker's only way out is to distinguish between the actual world (now understood as the way things are) and the world (understood as Lewis and his surroundings). But this, McGinn suggests, is an unwelcome move, precisely because the appeal of moderate realism derives, in large part, from its contrast between merely possible worlds as ways things might have been and the actual world as a comprehensive collection of bits and pieces. This is perhaps a fair point against Stalnaker's actual position, but it is worth noticing that there is space for a somewhat different version of moderate realism which escapes the objection.

The difficulty is an immediate consequence of Stalnaker's rejection of thesis (2). But thesis (2), as he understands it, is really two quite independent theses:

(2*a*) Other possible worlds are things of the same sort as the actual world
(2*b*) The actual world is David Lewis and his surroundings

Both theses are essential to Lewis's full-blooded realism; in consequence, the rejection of either separately defines a more moderate position. Retention of (2*b*) coupled with rejection of (2*a*) – the option Stalnaker actually plumps for – is what leads to the problem just discussed. If, instead, we retain (2*a*) but reject (2*b*), the problem does not arise. On this position, every possible world is a way things might have been, the actual world being just that one among them which is the way things happen to be. Moderate realism of this kind can, in contrast with Lewis's, claim the support (for what that is worth) of his paraphrase argument for possible worlds. If ways things might have been are thought of as determined by maximally consistent sets of propositions, realism of this kind enjoins a broadly Tractarian conception of possible worlds and perhaps coincides with a position defended by some contributors to the debate (cf. Adams, 1974; Hintikka, 1969; Plantinga, 1974, ch. 4).

It is arguably at a considerable advantage over Lewisian realism, since it is clear enough how, on this view, atomic propositions are guaranteed determinate truth-values at each possible world, whereas that is anything but clear, if the actual world consists merely of Lewis

and his surroundings, and other worlds are likewise conceived as comprehensive collections of similar bits and pieces, at least unless proposition-like entities are smuggled in under the somewhat opaque heading of 'surroundings.' The view will, of course, have little appeal for those, such as Lewis and Stalnaker, who think propositions are best analyzed as sets of possible worlds; but it is more than a little questionable whether that can withstand scrutiny.[30] It should, on the other hand, be congenial to those philosophers who, in broadly Fregean tradition, hold ontological questions to be best conceived as questions about truth and logical form. (See Chapter 20, REALISM AND ITS OPPOSITIONS, §1.)

A more recent and, in some respects, more radical suggestion, aimed at securing the presumed advantages of construing modal operators as quantifiers over possible worlds at bargain price, is what Gideon Rosen calls *modal fictionalism*. On this proposal[31] we should prefix possible-world paraphrases of modal statements with a non-factive operator which suspends commitment to the possible worlds over which the statement to which it is prefixed quantifies. Much as prefixing 'According to Genesis 19:26' to 'Lot's wife was turned into a pillar of salt' produces a compound statement which we may assert and believe without committing ourselves to the actual occurrence of the saline transformation of which the component purports to speak, so – the fictionalist proposes – we may seal off the unwanted ontological commitment carried by $\exists w \mathrm{P}^{*} w$ (the possible-worlds version of $\Diamond p$) by prefixing it with 'According to PW,…,' where 'PW' denotes some suitable version of possible-worlds theory (such as Lewis's).

It is far from clear that embedding possible-world paraphrases of modal statements really does leave us with a theory which enjoys all the supposed advantages of Lewisian realism.[32] Even prescinding from worries on that score, however, the proposal appears to fall foul of a simple dilemma. Observe first that regardless of whether he accepts the conditional 'If PW were true, it would be true that *A*' as a fully adequate explanation of what is meant by 'According to PW, *A*,' the fictionalist can hardly deny that each entails the other; or, if he does, then pending an explanation of what he does mean by the latter there is no theory to discuss. The whole point of fictionalism is, of course, to keep open the option of accepting fictionalized versions of quantifications over possible worlds, whilst rejecting the modal realist's ontology, or at least going agnostic. For simplicity, let's suppose the fictionalist wants to go atheist. So he thinks that PW is false. So, does he think it is contingently false, or that it is necessarily false? If the latter, then he runs into trouble immediately – whatever modal statement *p* is, his replacement for its possible-world translation is going to be vacuously true, simply by virtue of the necessary falsehood of its antecedent. If, instead, he opts for the view that PW, though false, is no worse than contingently so, he must hold that PW might be (or might have been) true. The problem now is to see how this claim, that possibly PW is true, is to be understood. Obviously it cannot be understood in the fictionalist's preferred manner; replacing it by a fictionalist ersatz of the usual kind gives us: 'If PW were true, then there would be a world at which PW is true.' Since this is a direct consequence of 'If PW were true at the actual world, it would be true at the actual world,' it would be true, as would the latter, even if PW was necessarily false. But then the paraphrase can scarcely be held to capture the content of the claim that PW is possibly true. However, if there is some other way to understand this particular modal claim, equally free of commitment to suspect ontology, then it is unclear why modal claims quite generally should not be understood in that way, with the upshot that fictionalization loses its point.[33]

Both of the proposed alternatives to Lewis's realism just discussed make one very important concession to it, namely that ordinary modal idioms are best understood in terms of

quantification over possible worlds (though as noted, the fictionalist, at least, appears obliged to recognize a use of some modal idiom which cannot be reduced to such quantification). It is far from clear, however, that so much should be conceded. Several writers have advocated semantical accounts of modality on which modal adverbs are treated as what surface syntactical form suggests they are, that is, a species of sentential operator. These accounts typically take the form of showing how a truth-theory for an object-language including modal operators may be constructed in a metalanguage which itself contains either those same operators (in which case the relevant clauses can be homophonic) or counterparts which are direct translations of the object-language operators (cf. Peacocke, 1978; Davies, 1981, part III; Forbes, 1985; 1989). Lewis has not, as far as I know, explicitly criticized this approach, but it is not hard to guess at his most likely response. In *Counterfactuals* (Lewis, 1973) he asks: 'If our modal idioms are not quantifiers over possible worlds, what else are they?' and takes it, apparently, that there are just three significantly different alternatives: (1) to take them as unanalyzed primitives, (2) to interpret them as metalinguistic predicates, analyzable in terms of consistency (for example, 'Possibly p' means that p is a consistent sentence), and finally (3) to take them as quantifiers, but ones ranging over a domain of some kind of Ersätze, such as maximally consistent sets of sentences. Options (2) and (3) fall, he thinks, to the objection that they simply give incorrect results unless 'consistent' is understood as 'possibly true,' in which case the theory is circular; (1) he dismisses as "not an alternative theory at all, but an abstinence from theorizing" (Lewis, 1973, p. 85). The force of the circularity objection may be doubted, since we have no right to expect philosophically interesting notions always to admit of fully reductive analyses; and there is, in any case, some question whether Lewis is well placed to press it, since he appears himself to have to rely upon an unanalyzed notion of possible world.[34] He might rejoin that whilst 'possible world' is indeed a primitive for him, the notion receives elucidation via its deployment in his theory in a manner akin to that in which it is commonly held that theoretical terms in natural scientific theories do. But this is unconvincing. Whatever its technical utility, possible-worlds theory holds out little promise of illuminating answers to the philosophical questions about necessity and possibility which exercise us: What is the source or ground of possibility, and how in general do we get to know about it?

3.4 Objections to Lewisian Realism

A viable alternative to Lewis's realism, perhaps along one of the lines we have considered, which can match whatever explanatory virtues may legitimately be claimed for it would undermine Lewis's main argument for his position; but that would not, of course, constitute a direct argument against it (although some, keen to wield Occam's Razor, might see it as a strong indirect argument). There can be no doubt, however, that the principal spur to attempts to develop such an alternative has been the conviction that full-blooded realism should be avoided if at all possible. The sources of this conviction are many and various. We cannot review them all here, but will conclude this part of our discussion with an examination of one particular line of objection which several thinkers regard as the most important direct argument against Lewis's view, and which, together with Lewis's response to it, broaches questions bearing on wider issues in the philosophy of modality. This is the argument from epistemology. If, as in Lewis's view, we stand in no sort of causal, or other natural, relations with possible worlds other than our own, or with their inhabitants, how can we possibly have knowledge, or even well-grounded beliefs, about them? And if the

truth-conditions of ordinary modal propositions are as the modal realist maintains – that is, if they are most perspicuously set forth by paraphrasing them as quantifications over possible worlds – is not the disastrous effect of modal realism that modal knowledge and well-grounded modal belief are rendered impossible?

If it were, in every case, a necessary condition for X to know that *p* that X's belief that *p* should be caused by (or stand in some other suitably causal relation with) the fact that *p*, then the objection would be decisive. It thus appears that the modal realist must deny that knowledge is invariably subject to such a causal constraint.[35] And his position will be the more plausible if he can furnish independent ground for doing so. Lewis contends that mathematical knowledge affords the desired precedent for rejecting a fully general causal requirement. The causal epistemologist's objection to modal realism runs parallel, he claims, to Benacerraf's celebrated dilemma for mathematical knowledge. This rests upon the idea that there is a head-on collision between the demands of a broadly causal conception of knowledge on the one hand, and on the other, any account of the truth-conditions of mathematical propositions which has them speaking of causally inert mathematical objects (numbers, sets, and so on). It is clear, Lewis thinks, how we should respond to Benacerraf's problem: "our knowledge of mathematics is ever so much more secure than our knowledge of the epistemology that seeks to cast doubt on mathematics," so it is the latter which must go – "Causal accounts of knowledge are all very well in their place, but if they are put forward as general theories, then mathematics refutes them" (Lewis, 1986, p. 109).

It would be a perfectly fair objection to Lewis's response as it stands that it simply conflates mathematics with a certain philosophical account of it, according to which the surface syntax of ordinary mathematical statements is to be accepted at face value, with the consequence that numerals and many other mathematical expressions are to be regarded as genuine singular terms, having reference among abstract objects of various kinds. It is the latter, not the former, which is (supposedly) put in doubt by the causal epistemologist's objection. But waive that: even if the case were soundly made that mathematical knowledge should not be seen as demanding causal connections between knowers and what they know, it might still be objected that there is a crucial difference between this case and modality as the modal realist conceives it. Lewis recognizes this: the mathematical objects of which we have knowledge, for all our lack of causal acquaintance with them, are abstract, whereas other possible worlds and their occupants are, as Lewis conceives them, no less concrete than this world and its occupants. There is, then, space for the counter that it is precisely and only because mathematical entities are abstract that we should not expect mathematical knowledge to satisfy a causal condition, so that the suggested precedent is not enough to get the modal realist off the epistemological hook (Lewis, 1986, p. 110). Lewis's response is, in effect, that this mis-identifies what it is about the mathematical case that warrants suspension of causal requirements on knowledge:

> causal acquaintance is required for some sorts of knowledge but not for others. However, the department of knowledge that requires causal acquaintance is not demarcated by its concrete subject matter. It is demarcated instead by its contingency.… [Perception and] other channels of causal acquaintance set up patterns of causal dependence whereby we can know what is going on around us. But nothing can depend counterfactually on non-contingent matters.

Among the non-contingent matters are what mathematical objects there are, and likewise, what possibilities there are. And this is why the imposition of a causal constraint is inappropriate in both cases alike (Lewis, 1986, p. 111).

Even if Lewis's claim about how the area within which imposition of a causal constraint upon knowledge should be demarcated is correct, and even if, further, his main argument for realism from its alleged explanatory advantages is successful in its own terms, there would still be room to question whether he has done enough to see off the epistemological challenge. He would have done so only if the latter argument justifies us in taking claims about what possible worlds there are, and what they are like, to report non-contingent matters, and it is far from evident that it does so. It might, to be sure, be held that statements about what is necessary or what is possible (at least, when broadly logical modality is in question) are themselves, if true at all, necessarily so (as in the modal logic S5). But again, that is not enough for Lewis; the issue concerns the modal status not of ordinary modal claims themselves, but of their construals as quantifications over possible worlds as Lewis understands them. Even if it is allowed that an argument to best explanation may warrant taking such claims to be true, it is hard to see how it could justify us in taking them to be *necessarily* true.

Lewis's claim that what marks off the area within which knowledge should satisfy a causal constraint is not concreteness but contingency is plausible, and coheres with a plausible explanation why the line should be drawn where Lewis proposes to draw it. Satisfaction of a causal constraint is to be looked for just when there can, but need not, be a significant covariation between our beliefs and the facts which confer truth or falsehood upon them. Since non-contingent matters are precisely ones which could not have been otherwise, any counterfactual conditional hypothesizing the falsehood of a non-contingent (that is, a necessary) truth must be vacuously true, whatever its consequent says, for example about what we would then have believed. But then there can be no significant covariation between the non-contingent facts and our beliefs about them, so there is no sense in requiring that they stand in an appropriate causal relation.

Plausible it may be, but the claim is not beyond question. Indeed, if Kripke and those who follow him are right in their claim that there are necessities – such as that water is H_2O and that heat is mean kinetic energy – which can be known only *a posteriori*, then it appears to be mistaken. Such apparent counter-examples might be explained away by maintaining that a different notion of necessity is involved in them, but it is not easy to see independent grounds for supposing that to be so. A more plausible reaction would be to modify the principle of demarcation to something along these lines: causal constraints upon knowledge are inappropriate when we are concerned with necessities known, or knowable, *a priori*. But that suggests that the initial emphasis on non-contingency as such was misplaced, and that the important contrast here is not between necessity and contingency, but between *a priori* and *a posteriori* knowledge. That is, it is a truth's being known independently of experience (however that notoriously problematic notion is precisely to be characterized) that renders inappropriate the demand that our knowledge of it should satisfy a causal constraint.[36] It might be suggested that from this improved perspective, Lewis's arguments from explanatory virtue would, other things being equal,[37] be better suited to their purpose than our preceding remarks suggest. There is, it might be claimed, no evident reason in principle why the explanatory virtues of a realistic attitude towards possible worlds should not lie in its underpinning independently plausible analyses of counterfactuals, propositions, and other *a priori* matters. But this sanguine response overlooks a crucial distinction: it may be that there could be a successful argument from explanatory virtue for something which is in fact knowable *a priori*; but it obviously fails to follow from this, and is anything but clearly true, that such an argument could itself provide us with *a priori* knowledge.

We have thus far followed Lewis's own discussion of the charge that his position is epistemologically bankrupt largely in presupposing that it will be based upon a causal constraint on knowledge. But it may well seem that his position is in epistemological trouble even if a causalist – or more generally, naturalistic – view of knowledge is *not* assumed. Other worlds, in Lewis's view, are composed of concrete entities possessed of properties and standing in relations to one another of the same general kind as the concrete entities in our world (the actual world). If another world contains knowing subjects anything like us, these subjects know of the doings and undergoings of things in their world much as we know of such things in ours. But because of their utter causal isolation from us, we cannot possibly know of those other worldly goings-on in anything like the ways they are supposed to do, and we must be supposed to know of them in some radically different way. But then whatever account may be proposed, it seems that a yawning chasm is bound to open up between the truth-conditions of ordinary modal statements (as Lewis conceives them) and our knowledge; nothing in the character of our knowledge could in any discernible way reflect the nature of the states of affairs which confer truth upon the propositions known.[38]

It might be replied that the objection misdescribes the propositional objects of our modal knowledge as they are best conceived on Lewis's view. There would indeed be a serious, and perhaps insurmountable, difficulty if our modal knowledge had to consist, or be grounded at a fundamental level, in knowledge of the doings and undergoings of particular identifiable objects existing in other worlds – at least on the plausible assumptions that such knowledge would require identifying reference to, or thought of, those objects, and that, in case of concrete objects, no such identifying thought is possible that does not depend, ultimately, on the obtaining of causal or other natural relations between thinker and object. But Lewis's view need not take this shape: what we know, when we know that possibly p or that necessarily q, is, in his view, a general proposition: that there is a world having such-and-such a character, or that all worlds satisfy a certain general description. This would be no help, of course, if our knowledge of such general truths had to be grounded in anterior knowledge of truths concerning particular worlds, as would be so if our knowledge that there is a world in which things are thus and so had to derive by existential generalization from the knowledge that things are that way in w_{17}, and our knowledge that in all worlds such-and-such had to be obtained by (ordinary) inductive inference from knowledge of how things are in some finite selection of worlds. It follows immediately that our modal knowledge, as Lewis conceives it, cannot be like our knowledge that there are cities in the United Kingdom with more than two million inhabitants, or other-worlders' knowledge of similar truths concerning their own world. But this is a point he might readily accept: there are other instances – which for present purposes may be regarded as uncontentious – in which our knowledge of general truths is not of that kind. The obvious examples are afforded by mathematics where, on a classical view at least, we may come to know general truths of both kinds by non-constructive methods, such as proof by *reductio ad absurdum*. The parallel with mathematical knowledge is, once again, one to which Lewis himself appeals in this context:

> In the mathematical case, … we come by our opinions largely by reasoning from general principles that we already accept; sometimes in a precise and rigorous way, sometimes in a more informal fashion, as when we reject arbitrary-seeming limits on the plenitude of the mathematical universe. I suppose the answer in the modal case is similar. I think our everyday modal opinions are, in large measure, consequences of the principle of recombination.[39] … One could

> imagine reasoning rigorously from a precise formulation of it, but in fact our reasoning is more likely to take the form of imaginative experiments. We try to think how duplicates of things already accepted as possible … might be arranged to fit the description of an alleged possibility. Having imagined various arrangements – not in complete detail, of course – we consider how they might aptly be described. (Lewis, 1986, pp. 113–114)

As a rough account of the phenomenology of ordinary modal thinking, this is scarcely open to dispute. But anyone who was troubled by the appearance of an uncomfortable gulf between, on the one side, any credible story about how we might get to know, or justifiably believe, propositions about what is possible or necessary, and on the other, their Lewisian truth-conditions, is liable to feel short-changed. The idea that the possibility of unicorns, for example, consists in there being some other possible world in which concrete horse-like entities have concrete horn-like appendages plays no essential part in the plausible part of Lewis's story about the exercise of imagination in tandem with his combinatorial principle. That is an account which anyone could offer, without commitment to realism about worlds, to which that realism is at best a gratuitous addition – at best: the case is arguably worse, since it is hard to see how the imaginative exercises which Lewis plausibly identifies as the source of our modal beliefs could possibly equip us with adequate reasons for those beliefs, if they really carried the ontological commitments he ascribes to them.

4 Modal Realism 2: The Non-cognitivist Challenge

We have not tried finally to resolve the issue of realism about possible worlds. For all the heat – and comparatively little light – it has generated, it is, in a fairly clear sense, something of a distraction from the leading questions identified in §1.4. The sense in which those questions presuppose a realist conception of modality appears not to be Lewis's, but a more modest one; what is at issue is, rather, the existence of a genuine class of truths essentially involving modal notions.[40] Or, to put the question another way, does a correct claim of the form 'Necessarily *p*' state a fact over and above the fact that *p*? Several recent writers, some of them more or less explicitly following a line of thought suggested by Wittgenstein's remarks on necessity in his *Remarks on the Foundations of Mathematics* (Wittgenstein, 1978), are united in advocating a negative answer to this question, whilst differing both in the considerations they adduce in its support and in their positive accounts of the role or function of necessitated judgments.[41] The principal consequence of the negative thesis is that, when we assert it to be necessary that *p*, we are not making any claim (over and above the plain claim that *p*) concerning which there arises any question about how we know it to be true. I shall, accordingly, use the term 'non-cognitivism' to denote the shared negative thesis.

Philosophical doctrines to the effect that the sentences belonging to a given region of discourse do not – syntactical and other appearances to the contrary notwithstanding – genuinely record or misrecord an appropriately corresponding range of facts have, of course, enjoyed a good deal of popularity, especially among philosophers of broadly empiricist sympathies. The obvious examples are sentences used to voice moral and aesthetic judgments. Faced with the more or less manifest inadequacy of attempts to construe such sentences as having naturalistically statable truth-conditions, and the apparently intractable problem of seeing how we might acquire moral or aesthetic knowledge by anything remotely resembling the methods with which – prescinding from radically skeptical

doubts – we feel comfortable in other territory, and unwilling to postulate a special realm of 'queer' facts and a suitably attuned, special mode of cognitive access to match, the option can readily appear attractive of supposing that moral and aesthetic talk is not aimed at describing or 'tracking' moral or aesthetic fact at all, but is best understood in some other way – as expressive of our own moral and aesthetic responses, say, or aimed at influencing the responses and actions of others. While it is clear that they could not be decisive, there is no doubt that similar considerations may play their part in motivating non-cognitivist thinking about modality.[42] We find the same distrust of irreducibly modal fact, and the position derives a good part of its attraction from the perceived inadequacy of attempts to provide a credible epistemology.

One line of thinking here focuses on the role of imagination in the genesis of modal opinion. Very often, we are moved to judge that things must be thus and so by the seeming unimaginability or inconceivability of the opposite. Confronted – to take what is arguably a fundamental kind of case – with what qualifies, by ordinary criteria, as a valid deductive inference, and finding ourselves unable to conceive how the premise could be true without the conclusion being so as well, we move, without much ado, to the belief that we are faced with a necessity. It is hardly surprising that non-cognitivists tend to look askance at this move, the relations between conceivability and possibility and their opposites having long been a matter of philosophical controversy. The non-cognitivist will grant the facts about what we can and can't imagine, and will assure us that he has not the slightest tendency to doubt that if the premise is true, the conclusion is true as well; but he will protest that he cannot see how that justifies a belief that the conclusion *must* be true, if the premise is. The limit of our imagination may well have a part to play in explaining our confidence in certain judgments, but it is just another fact about us: what reason is there to see in it a reliable indication not merely of their truth, but of their *necessary* truth?[43]

It seems clear that the cognitivist should concede right away that the step to necessity from our inability to conceive the opposite is problematic, if only because, in general, our being unable to do something may, so far, be properly explained either in terms of some limitation from which we perhaps contingently suffer, or in terms of impossibility inherent in the task, which our inability merely reflects. But that is not to concede that it is merely a confused and broken-backed attempt to inflate facts about our imaginative limitations into objective necessities. If the move is thought of as supplying our basis for thinking that there are such things as necessities at all, and is supposed to yield infallible access to them, then it surely is hopeless. But the cognitivist need not be committed to the implausible claim that we are equipped with an infallible method of detecting necessary truths. And he can insist that we should separate the question of our grounds for holding that there are necessary truths to be appreciated concerning some matter, from the question of how we may be justified in taking some particular proposition to be necessary. It is, for example, one thing to hold that if a number is prime, it is necessarily so, and another to hold, of some particular number, that it is necessarily prime; calculation may provide us with grounds for the latter opinion, but support of a quite different kind is needed for the former. If the cognitivist can sustain the distinction in general, and can make the case that there are necessities to be discerned, then he may argue that our inability to imagine things being otherwise can be taken as a fallible, defeasible ground for belief in the necessity in particular cases. These are, of course, very big 'ifs.' The present point is simply that, pending some demonstration that they cannot be discharged, considerations of the kind just rehearsed are bound to be inconclusive, and need not dislodge a determined cognitivist.[44]

Progress on the present issue seems unlikely in the absence of some general, agreed criteria for discriminating between cases in which statements concern some genuinely factual matter – in which correctness is properly seen as consisting in conformity with some range of independently constituted facts, which our opinions, as thereby expressed, may be regarded as in some sense tracking – and cases in which this is not so. A proposal very much to the purpose has been elaborated and defended by Crispin Wright, originally in his book on Wittgenstein's philosophy of mathematics (Wright, 1980), subsequently in his paper "Inventing logical necessity" (Wright, 1986), and in his Waynflete Lectures, *Truth and Objectivity* (Wright, 1992). The general idea underlying Wright's proposal is that statements of a given class are properly viewed as (mis)recording genuine matters of fact – or better, as potentially representing objects of knowledge or at least rationally justifiable opinion – only if there are, *a priori*, certain kinds of limitation on the possibility of intelligible but unresolved disagreement over their truth-values. (See Chapter 20, REALISM AND ITS OPPOSITIONS, §5.) Roughly, the thought is that where A is a statement of the kind in question, such disagreement is intelligible only if traceable to the operation of what can be regarded as a cognitive shortcoming in at least one of the parties to it. Besides omitting important refinements, this way of putting it is, of course, objectionably circular. Here is a more careful formulation, taken from the 1986 paper mentioned above:

> Statements of a certain class are apt for the expression of genuine matters of fact only if there are contexts – in which vagueness, or permissible differences in evidence thresholds, are not to the point – in which it is *a priori* that differences in opinion concerning one of the relevant statements can be fully explained only by disclosing … some material ignorance, error, or prejudice on the part of some or all of the protagonists.[45]

How, assuming its approximate correctness, does this criterion bear on our present question? It may at first appear that, in contrast, for example, with claims about what is funny or boring, where we are happy enough, on occasion, to write off differences of opinion as due simply to diverging tastes or interests, its application would favor the *cognitivist* about modal matters rather than his opponent. But, as Wright argues, matters are not so straightforward. Can we not conceive of a supremely cautious thinker who agrees with the rest of us on all relevant non-modal matters, but consistently balks at the point where we are disposed to judge something to be not just true in fact, but necessarily so? Suppose, then, that we find ourselves locked in apparently intractable disagreement with such a character over, say, the necessity of the conditional corresponding to some simple deductive inference, the correctness of which is agreed on both sides. Neither party, it seems, is under any misapprehension of the exact character of the formal transition in question, and both, it seems, are competent in the use of the logical vocabulary involved. Is it *a priori* that the persistence of our disagreement must, sooner or later, succumb to explanation which convicts him, or us, of some germane ignorance, error, or prejudicial assessment of the data? If not, then the non-cognitivist may claim victory. And since the intelligibility of such a disagreement appears not to depend on anything special to our chosen case, it appears that our cautious individual will be at the service of the non-cognitivist in all cases in which we are disposed to think ourselves confronted with a necessary truth (see Wright, 1986, pp. 202 ff.).

Actually, that final move, to a globally cautious stance on necessity, is very much open to question. If the argument which Wright himself develops against Quine's global empiricism (sketched in §2.3) is good, it establishes that there is no coherent epistemology which does

not acknowledge some judgments – centrally, judgments about what a given empirical theory plus logic entails – as *a priori*. And if that is so, the question arises of how the possibility of acquiring reason to believe such judgments *a priori* is to be explained. But then, as Wright himself puts it, "What better basis on which to found a satisfactory account of the possibility of arriving at certain truths by pure thought than on the notion that they hold true in all thinkable circumstances?" (Wright, 1989, p. 223) (that is, that the truths in question are necessary). It thus appears that there are, after all, reasons to doubt that a globally cautious stance is fully intelligible without supposing its adherent guilty of some cognitive shortcoming: either a failure to acknowledge an indispensable distinction between *a priori* and empirical methods of appraisal, or a failure to appreciate that this distinction issues from the necessity of some judgments and the contingency of others (cf. Wright, 1989, pp. 222–223).

Even if the case just sketched can be made secure, however, cognitivist celebrations would be premature. For on reflection, it seems that an effective case for non-cognitivism might be made *without* relying upon the dubious possibility of global caution. The anti-Quinean argument against the intelligibility of a globally cautious attitude does nothing to establish the unintelligibility of caution in any particular case. Why wouldn't the intelligibility of *local* caution, provided it can strike anywhere, suffice for the non-cognitivist's dialectical purposes? To put the thought another way, why shouldn't the destructive work that was to be done by a single, globally cautious thinker be distributed across a suitably large team of selectively cautious thinkers, each willing to affirm necessity in some cases, while remaining resolutely cautious in others?[46]

There are at least two reasons to doubt that this ingenious twist can accomplish what the non-cognitivist seeks. First, any puzzlement we may have felt about the intelligibility of the globally cautious attitude in relation to particular cases is liable to be compounded by the supposition that caution in selected cases is now coupled with normality (that is, absence of this peculiarly philosophical distaste for modalizing) in others. Are we to suppose, for example, that confronted with *some* valid inferences, our selectively Cautious Man has no compunction in agreeing that their premises necessitate their conclusions, and yet in other cases, simply refuses point-blank to do so, yet without having anything to say in explanation of his peculiar pattern of necessitated judgments? This begins to seem really unintelligible. If, on the other hand, there is some method in his apparent madness, there ought to be something to be said about the principles that inform his selective judgments; and we may suspect that when it is said, we shall be able to locate some material disagreement on other, non-modal matters. This difficulty is, clearly, special to the hypothesis of selective caution; the second, which could as well have been raised in relation to global caution, concerns whether cognitively blameless caution really is a fully intelligible attitude in every case. Suppose, for instance, that the Cautious Man is invited to pronounce upon an explicit formulation of the Law of Non-Contradiction, and responds thus:

> Hm. I am not sure that this is something that I can form a competent opinion about just by reflection. I cannot, I grant, recall any actual example of a statement which was true simultaneously with its negation. And I must confess to some difficulty when I try to be clear about how such a thing might occur. I suspect that it never does occur. Nevertheless, I do not see that this can be a matter for adjudication by *a priori* methods alone.[47]

As Wright observes, this stance is not intelligible. The reason why not is that its intelligibility would require the Cautious Man to believe that some further process is needed to establish

the falsehood of the negation of a proposition, even after the truth of the original proposition had been established – much as, having calculated the value of 10,987 + 3,733, a further process is needed to determine whether it is the same as or distinct from that of 174 × 80. And that, Wright points out, is absurd: "negation is given as a function on truth-value … To suppose that the truth-value of not-*p* may present an *a priori* open question when that of *p* has been settled is merely to display a failure to grasp that negation is, constitutively, a *truth-function*" (Wright, 1989, p. 230).

If this is right, then there are at least some cases in which caution is simply not an option at all, and is therefore unavailable as a means of enforcing a non-cognitivist view of necessity. Clearly crucial questions remain, which we cannot pursue here. So far, it may seem that the non-cognitivist has merely to give up one strategy, but that his position might yet be secured by other arguments. But the damage would be greater, if Wright's criterion could be taken as embodying an acceptable sufficient condition for a statement's enjoying genuinely factual status. For it would then be hard to see how non-cognitivism about a given range of necessitated judgments could be sustained without upholding the intelligibility of the cautious attitude (or something not materially different from it) in those cases. So two pressing questions are: Does Wright's criterion (or some near relative) embody an acceptable *necessary* condition of factuality? and: If so, does it also give an acceptable *sufficient* condition? A third question, which sets the agenda for anyone who hopes to defend affirmative answers to the last two, and is encouraged by the argument of the preceding paragraph, is: How far can considerations of the kind adduced in support of the claim that caution about the *a priori*/necessary status of the Law of Non-Contradiction is incoherent be duplicated in other putative cases of necessary truth?[48]

There have been other challenges to the more modest variety of realism under discussion in this section. Prominent among them is a dilemma by which Simon Blackburn seeks to undermine what he terms the 'truth-conditions approach' to modality (see Blackburn, 1986, pp. 120–121; or Blackburn, 1993, pp. 53–54). Suppose we ask: Why is it necessary that *p*? and receive the answer: Because *q*. Then either *q* will claim just that something *is* so, or that something *must* be so. But if *q* is true but might not have been, then it might not have been true that *p* either – *p*'s supposed necessity is 'undermined.' But if *q* is put forward as necessary, we have either a vicious circle or the start of a vicious regress. We have at best merely explained one necessity in terms of another. But we want to know why anything at all is necessary. Whether this simple dilemma is as effective as Blackburn hopes is a question I must leave the reader to ponder.[49,50]

Notes

1 The case for the indispensability of modal notions may, of course, be argued in other ways. In particular, Timothy Williamson makes a strong case that while most scientific theories are expressible in wholly non-modal language, their application, and in some case their proper interpretation, requires deploying modal notions. See Williamson (2016)

2 For the standard account of relative necessity, see Smiley (1963). Some serious problems for it are raised in Humberstone (1981) (see also Humberstone, 2004).

3 Since state of information varies from thinker to thinker and time to time, the precise import of such epistemic uses of modal idioms would involve also some relativity to context.

4 See, for example, McFetridge (1990, p. 136).

5 Note that this does not amount to assuming comparability with logical necessity and possibility.

6 This argument, and its further significance, are discussed in Hale (1996), and more recently in Hale (2013, §4.3).

7 Dummett (1978, p. 169). The original article appeared as Dummett (1959).

8 "Possibility and necessity" reprinted in Putnam (1983); the words appear on p. 51.

9 Cf. Putnam (1983, p. 53), where Putnam writes: "There is, however, a very different way in which one can try to save the subject of 'necessity' from Quine's attack. Quine, following the logical positivists, assumed that if there was any such thing as 'necessity' then it was either semantical (e.g., 'analyticity') or epistemic ('apriority'). To Saul Kripke is due the honor of introducing into the discussion a very different kind of necessity, an objective non-epistemic kind of necessity: metaphysical necessity. Or so he called it."
In fact, Kripke does not use the term 'metaphysical necessity' in Kripke (1980); but he does contrast the notion of necessity as a metaphysical one with the notions of analyticity and apriority, which he takes to be semantic and epistemological, respectively.

10 Putnam goes on to express some reservations about the strong claim that water is H_2O in all possible worlds, suggesting that "the 'essence' that physics discovers is better thought of as a sort of paradigm that other applications of the concept … must resemble than as a necessary and sufficient condition good in all possible worlds" (Putnam, 1983, p. 64): but he does not see this as undermining the response to Quine's attack.

11 Kripke's idea – his simple inferential model – is taken up and further developed in Hale (2013, §11.2 ff.).

12 It should perhaps be stressed that the point here is purely epistemological – that *a posteriori* knowledge of (metaphysical) necessities depends upon *a priori* knowledge of what would normally be taken (by anyone who has a use for the notion) to be conceptual necessities. It does not appear to require the claim that metaphysical necessity can be analyzed in terms of conceptual necessity – though of course, if such an analysis could be provided it would supply an independent reason against Putnam's proposal. Alan Sidelle (1989) tries to explain how, compatibly with the view that all necessity derives from conventions, there can be *a posteriori* necessities such as that water is H_2O. His general idea is that such metaphysical necessities are grounded in 'general principles of individuation' which record analytic, conventionally grounded necessities. An example of the latter would be 'If water has a certain chemical composition, it has that chemical composition necessarily.' Sidelle explicitly distances himself from the logical empiricists' thesis that all necessity is analytic, and appears not to view his general principles as providing the basis for an analysis of metaphysical in terms of analytic necessity. Just as well, since the necessity operator governing the consequent in such principles can hardly be regarded as expressing analytic necessity. We cannot here pursue the question of whether he succeeds in developing a viable alternative which does not simply boil down to the Kripkean explanation of how *a posteriori* necessities may be known. If he does, that would provide a third reason why Quine's attack on *a priori* necessity cannot be finessed by Putnam's proposal.

13 For a fuller discussion of the issue, see Chapter 23, ANALYTICITY. A third, and very interesting, attempt to prove that we cannot dispense with the idea that some statements are logically necessary may be found in McFetridge (1990, pp. 153 ff.). McFetridge's argument is discussed and developed in Hale (1999), which is in turn criticized in Ahmed (2000), to which Hale (2000a) replies. I give a somewhat simplified, and perhaps more digestible discussion in Hale (2013, ch. 2).

14 Roughly, strike the best balance between minimizing 'recalcitrance' (clashes between our total set of accepted statements and experience) and maximizing overall simplicity and economy to theory.

15 Note that this is *not* the option of tinkering with the underlying logic.

16 Perhaps with the exception of the special case of *a posteriori* necessities concerning natural kinds, etc.

17 Cf. Wright (1980, ch. 23; 1986), but see also Wright (1989), where Wright acknowledges some significant limitations on the scope for non-cognitivism. See below, §4, especially pp. 827 ff.

18 Or equivalent formulae, such as $\Diamond\Diamond A \rightarrow \Diamond A$ and $\Diamond\Box A \rightarrow \Box A$.

19 We would not expect the converse of the B axiom, i.e., $\Box\Diamond A \rightarrow A$ to be a theorem of any significant modal system. If it were added to S5 as an axiom, we could prove the equivalence of $\Box A$ and $\Diamond A$ with plain A, and so with each other, so that the modal system would effectively collapse down to its underlying non-modal logic.

20 For a classic discussion, see Lemmon (1959), which argues that no single system is correct – rather, which system is correct depends upon how the notion of necessity is understood. But even when a more or less definite kind of necessity is understood, there may be serious disagreement. For example, some have argued that the right modal logic for metaphysical necessity is S5; but others have argued that even the weaker system S4 is too strong. See Williamson (2013, ch. 3), Hale (2013, §5.4); but contrast Salmon (1989) and further work there cited.

21 The idea is often thought to come from Leibniz. But as far as I know, although Leibniz famously held that God made actual the best of all possible worlds, he nowhere explicitly proposes to explain necessary truth as truth at or in all possible worlds.

22 Cf. Wittgenstein (1961, §1.1): "The world is the totality of facts, not of things."

23 The lack of fit between the paraphrase argument and Lewis's prevailing conception of worlds is noted in Haack (1977) and Richards (1975).

24 Lewis (1973, p. 86). The objection has no force against the Tractarian conception of possible worlds which best fits the paraphrase argument.

25 For the details of Lewis's account of the truth-conditions of counterfactuals, see Lewis (1973, ch. 1). For a detailed account of what he takes to be the distinctive explanatory pay-offs of his modal realism, see Lewis (1986, ch. 1).

26 According to this analysis, each world is, from the standpoint of its inhabitants, the actual world, and other worlds are merely possible, much as 'here,' as employed by a given speaker, denotes where she is, and any other place is, for her there, elsewhere. Just as none of us is tempted to think that here – where we are currently – is the only place, so no one should be tempted to suppose that this world – the actual world, from our point of view – is the actual world. Being real is not to be identified with being actual.

27 It is fairly clear that Stalnaker takes properties to be abstract entities. Actually he says, rather unfortunately in my view, that they are abstract objects. This might suggest that he thinks that the actual world is, by contrast, a concrete object. Perhaps this way of making the contrast between the actual world and merely possible worlds could be sustained, if we thought of David Lewis and his surroundings as some kind of physical aggregate, as opposed to a set or collection. But it seems best to understand Stalnaker's view as being that possible worlds are properties, each of which might have been instantiated by the actual world, but only one of which is – the way things are.

28 Stalnaker's noted tendency to regard properties as (abstract) objects may partly account for his failure to remark on this point. It is not quite clear that possible-world quantifiers would have to be higher than second-order. Certainly if W is a predicate specifying a way things might have been, there is no evident reason why W should not involve quantification over first-level properties (i.e., properties of individuals), over properties of such properties, and so on, up to quantifications of arbitrary finite order. But this does nothing to prevent W from expressing a first-level property (which the actual world either has or lacks), any more than the higher-order quantification embedded in 'has all the qualities of a great general' prevents it from standing for a first-level property.

29 See McGinn (1981, p. 159). McGinn raises other objections which will not be discussed here.

30 For defense of analyses of propositions as sets of worlds, see Lewis (1986, ch.1) and Stalnaker (1984). The preceding sentence in the text indicates one potentially lethal objection. The issue cannot be pursued here.

31 Cf. Rosen (1990). Rosen's enthusiasm for the position has since been somewhat dampened – see Rosen (1993), and Brock (1993) for a similar problem. Others have been unpersuaded by the objection – see Menzies and Pettit (1994) and Noonan (1994).

32 For argument to the contrary, see Divers (1995).

33 This difficulty is elaborated in Hale (1995b). Rosen seeks to meet it in Rosen (1995), to which Hale (1995a) replies. The dilemma is further discussed by Daniel Nolan (2016), who finds it unconvincing.

34 Added 2014: This strikes me now as somewhat unfair, given that Lewis (1986) at least purports to provide a reductive explanation of what possible worlds are in non-modal terms. Whether the attempted reduction is satisfactory is, of course, a further question. For some discussion, see Divers (2002, ch. 7), Divers and Melia (2002), and, for a somewhat opposed view, Hale (2013, ch. 3)

35 This is a very strong assumption. There is, plausibly, reason enough to reject it, independently of the case in hand – it is, for example, difficult to see how it could fail to preclude the possibility of knowledge of, or justified belief in, perfectly ordinary empirical general truths (see, e.g., Hale, 1987, ch. 4, esp. pp. 92–101). But the difficulty can be put, as Hartry Field has observed in connection with mathematical knowledge (see Field, 1989, pp. 230–239), without relying on any such contentious claim about the analysis of the notion of knowledge. The modal realist should agree that we enjoy a significant degree of reliability in our modal beliefs – that we are fairly good at forming beliefs which accord well with the modal facts. On his view, this consists in our being good at forming beliefs which accord well with the facts about what (other) possible worlds there are, and the character of those worlds. If we are thus reliable, this is something which surely calls for explanation. The objection may then be put without appeal to a specifically causal analysis of knowledge or reasonable belief; given our lack of causal or other natural relations to other possible worlds, it is hard to see how any satisfactory explanation of the sort required could possibly be constructed.

36 McGinn (1976) actually proposes an explication of the *a priori*/*a posteriori* distinction in just such terms. It is important here not to forget that the kind of causal constraint in question is a strong one, to the effect that there must be a suitable causal relation between the putative knower's belief that *p* and the fact that *p*. Thus rejecting it for a given range of cases is not setting one's face against the possibility of any kind of causal or naturalistic account of knowledge whatever.

37 Whether other things are equal – and, crucially, whether Lewisian realism really does enjoy a distinctive advantage in such matters – is, of course, a further question.

38 McGinn (1981, pp. 153–158) develops an objection along these lines.

39 "Roughly, the principle is that anything can coexist with anything else, at least provided they occupy distinct spatio-temporal positions. Likewise, anything can fail to coexist with anything else" (Lewis, 1986, p. 88). Lewis is not committed to this initial formulation – it requires, in his view, a proviso to the effect that recombinations must be consistent with 'some possible size and shape of spacetime' (Lewis, 1986, p. 90).

40 This conception of modal realism is emphasized in, for example, McGinn (1981).

41 Thus Blackburn (Blackburn, 1984, ch. 6.5) depicts modal judgments as expressive of attitude, or something like an attitude; Wright (1980, ch. 23) explores the idea that they record decisions; while Craig (1985) speaks in terms of endorsement of a policy. Wright's position undergoes some shift in Wright (1989), where he grants that necessitated judgments are (at least minimally) truth-apt.

42 As they have in fact played. Blackburn's dilemma at (1986, p. 120) appears to be underpinned by unwillingness to accept irreducible modal facts; the epistemological motive is to the fore in Craig (1985).

43 Cf. Craig (1985, p. 93): "The limit of his imagination … is still just another fact about him, and he sees no reason to take it as a guide to what must of necessity be the case"; and Wright (1980, p. 453): "… you are asking me to affirm that whenever exactly the specified sequence of transformations is correctly followed through on exactly the specified basis, we are bound to achieve this (sort of) result – that no other outcome is possible provided the blueprint is correctly implemented. And that very strong claim, I feel, I am not entitled to make." It is true that Craig and Wright are here describing the position of the Cautious Man – an invention of Wright's, designed to unsettle cognitivists – not that of the non-cognitivist as such. But on the material point – that

the uncontroversial facts cannot rationally warrant a judgment of necessity – their views coincide; they differ, principally, in that while the Cautious Man thinks this obliges him to refrain from making necessitated judgments altogether, the non-cognitivist takes it to show that such judgments are not recognitional, but are to be understood as expressing or endorsing a policy or decision of some sort.

44 See Yablo (1993) for an excellent discussion of the standard objections to treating conceivability as a ground for possibility.

45 Wright (1986, pp. 199–200). For an earlier version of Wright's criterion, see Wright (1980, pp. 448–449). The reference to ignorance and error in the *explanans* should be understood, of course, as relating to ignorance or error about matters recorded in statements lying outside the disputed class – else the criterion would be hopelessly circular. Note that these may include facts about the meanings of relevant expressions.

46 Wright suggests, not a series of selectively cautious thinkers as here, but a single Eccentric Modalizer (cf. Wright, 1989, pp. 225–228). But so far as I can see, the net effect is the same, as are the problems with the suggestion.

47 This formulation is taken from Wright (1989, p. 229), as is the argument that follows.

48 For some – dare I say it? – cautious steps towards an answer to this third question, see the closing nine pages of Wright (1989).

49 For some argument to show that it isn't, see Hale (2002b), Cameron (2010), and Hale (2013, pp. 91–97).

50 Thanks to Crispin Wright and Nick Zangwill for very helpful comments on the original chapter, and to Jess Leech for helpful suggestions about updating it.

References

Adams, R. M. 1974. "Theories of actuality." *Noûs*, 8(3): 211–231.

Ahmed, A. 2000. "Hale on some arguments for the necessity of necessity." *Mind*, 109(433): 81–91.

Ayer, A. J. 1946. *Language, Truth and Logic*, 2nd edn. London: Gollancz.

Blackburn, S. 1984. *Spreading the Word*. Oxford: Clarendon Press.

Blackburn, S. 1986. "Morals and modals." In *Fact, Science and Morality: Essays on A. J Ayer's Language, Truth and Logic*, edited by G. McDonald and C. Wright, pp. 119–141. Oxford: Blackwell.

Blackburn, S. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Brock, S. 1993. "Modal fictionalism: a reply to Rosen." *Mind*, 102(405): 147–150.

Cameron, R. 2010. "On the source of necessity." In *Modality: Metaphysics, Logic, and Epistemology*, edited by B. Hale and A. Hoffmann, pp. 137–152. Oxford: Oxford University Press.

Craig, E. 1985. "Arithmetic and fact." In *Essays in Analysis*, edited by I. Hacking, pp. 89–112. Cambridge: Cambridge University Press.

Davies, M. 1981. *Meaning, Quantification and Necessity*. London: Routledge and Kegan Paul.

Divers, J. 1995. "Modal fictionalism cannot deliver possible world semantics." *Analysis*, 55(2): 81–89.

Divers, J. 2002. *Possible Worlds*. London and New York: Routledge.

Divers, J., and J. Melia. 2002. "The analytic limit of genuine modal realism." *Mind*, 111(441): 15–36.

Dummett, M. 1959. "Wittgenstein's philosophy of mathematics." *Philosophical Review*, 68(3): 324–348.

Dummett, M. 1978. *Truth and Other Enigmas*. London: Duckworth.

Field, H. 1989. *Realism, Mathematics and Modality*. Oxford: Blackwell.

Forbes, G. 1985. *The Metaphysics of Modality*. Oxford: Oxford: Clarendon Press.

Forbes. G. 1989. *Languages of Possibility*. Oxford: Blackwell.

Haack, S. 1977. "Lewis's ontological slum." *Revue of Metaphysics*, 30(3): 415–429.

Hale, B. 1987. *Abstract Objects*. Oxford: Blackwell.

Hale, B. 1995a. "Modal fictionalism – a simple dilemma." *Analysis*, 55(2): 63–67.

Hale, B. 1995b. "A desperate fix." *Analysis*, 55(2): 74–81.

Hale, B. 1996. "Absolute necessities." *Philosophical Perspectives*, 10: 93–117.
Hale, B. 1999. "On some arguments for the necessity of necessity." *Mind*, 108(429): 23–52.
Hale, B. 2000a. "Reply to Ahmed." *Mind*, 109(433): 93–96.
Hale, B. 2002b. "The source of necessity." *Noûs Supplement: Philosophical Perspectives*, 16: 299–319.
Hale, B. 2013. *Necessary Beings: An Essay on Ontology, Modality, and the Relations Between Them*. Oxford: Oxford University Press.
Hintikka, K. J. 1969. *Models for Modalities*. Dordrecht, Netherlands: Reidel.
Humberstone, I. L. 1981. "Relative necessity revisited." *Reports on Mathematical Logic*, 13: 33–42.
Humberstone, I. L. 2004. "Two-dimensional adventures." *Philosophical Studies*, 118(1–2): 17–65.
Hume, D. 1960. *A Treatise of Human Nature*. Oxford: Clarendon Press.
Kripke, S. 1971. "Identity and necessity." In *Identity and Individuation*, edited by M. K. Munitz, pp. 135–164. New York: New York University Press.
Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.
Lemmon, E. J. 1959. "Is there only one correct system of modal logic?" *Proceedings of the Aristotelian Society*, suppl. vol. 33: 23–40.
Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.
Linsky, B., and E. N. Zalta. 1994. "In defense of the simplest quantified modal logic." *Philosophical Perspectives*, 8: 431–458.
Linsky, B., and E. N. Zalta. 1996. "In defense of the contingently nonconcrete." *Philosophical Studies*, 84(2–3): 283–294.
Lowe, E. J. 2012. "What is the source of our knowledge of modal truths?" *Mind*, 121(484): 919–950.
McFetridge, I. 1990. "Logical necessity: some issues." In *Logical Necessity and Other Essays*, edited by J. Haldane and R. Scruton, pp. 135–154. London: Aristotelian Society.
McGinn, C. 1976. "A priori and a posteriori knowledge." *Proceedings of the Aristotelian Society*, 76: 195–208.
McGinn, C. 1981. "Modal reality." In *Reduction, Time, and Reality*, edited by R. Healey, pp. 143–87. Cambridge: Cambridge University Press.
Menzies, P., and P. Pettit. 1994. "In defence of fictionalism about possible worlds." *Analysis*, 54(1): 27–36.
Nolan, D. 2016. "Modal fictionalism." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/entries/fictionalism-modal/ (accessed August 27, 2016).
Noonan, H. 1994. "In defence of the letter of fictionalism." *Analysis*, 54(3): 133–139.
Peacocke, C. 1978. "Necessity and truth-theories." *Journal of Philosophical Logic*, 7(1): 473–500.
Peacocke, C. 1999. *Being Known*. Oxford: Clarendon Press.
Plantinga, A. 1974. *The Nature of Necessity*. Oxford: Clarendon Press.
Putnam, H. 1983. *Realism and Reason: Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press.
Richards, T. 1975. "The worlds of David Lewis." *Australasian Journal of Philosophy*, 53: 105–118.
Rosen, G. 1990. "Modal fictionalism." *Mind*, 99(395): 327–354.
Rosen, G. 1993. "A problem for fictionalism about possible worlds." *Analysis*, 53(2): 71–81.
Rosen, G. 1995. "Modal fictionalism fixed." *Analysis*, 55(2): 67–73.
Salmon, N. 1989. "The logic of what might have been." *Philosophical Review*, 98(1): 3–34.
Sidelle, A. 1989. *Necessity, Essence, and Individuation*. Ithaca, New York: Cornell University Press.
Smiley, T. 1963. "Relative necessity." *Journal of Symbolic Logic*, 28: 113–134.
Stalnaker, R. 1976. "Possible worlds." *Noûs*, 10(1): 65–75.
Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, R. 2003. *Ways a World Might Be*. Oxford: Clarendon Press.
Stalnaker, R. 2012. *Mere Possibilities*. Princeton, NJ and Oxford: Princeton University Press.
Williamson, T. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press.
Williamson, T. 2016. "Modal science." *Canadian Journal of Philosophy*, 46(4–5): 453–492. Also in *Williamson on Modality*, edited by M. McCullagh and J. Yli-Vakkuri. London: Routledge, 2017.
Wittgenstein, L. 1961. *Tractatus Logico-Philosophicus*. London and New York: Routledge and Kegan Paul.

Wittgenstein, L. 1978. *Remarks on the Foundations of Mathematics*, 3rd edn. Oxford: Blackwell.
Wright, C. 1980. *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.
Wright, C. 1986. "Inventing logical necessity." In *Language, Mind, and Logic*, edited by J. Butterfield, pp. 187–209. Cambridge: Cambridge University Press.
Wright, C. 1989. "Necessity, caution and skepticism." *Aristotelian Society*, suppl. vol. 63: 203–238.
Wright, C. 1992. *Truth and Objectivity*. Cambridge, MA, and London: Harvard University Press.
Yablo, S. 1993. "Is conceivability a guide to possibility?" *Philosophy and henomenological Research*, 53(1): 2–42.

Postscript

BOB HALE

The metaphysics, epistemology, and logic of modality are all areas in which much good and interesting work has been done in the two decades since this chapter was first put together. It will not be possible to comment on it more than cursorily. Some of it is the subject of newly added chapters (see Chapter 32, RELATIVISM ABOUT EPISTEMIC MODALS; Chapter 37, TWO-DIMENSIONAL SEMANTICS). This postscript discusses what the present writer views as some of the most important developments, as well as plugging some of the gaps left open in the original chapter.

The Source of Necessity and Possibility

If the various challenges to modal realism in our second sense (see p. 820) can be met, Dummett's questions (see p. 820) must be faced. His first question *can* be interpreted in a reductive spirit, as asking for an explanation of modal notions, or modal facts, in non-modal terms. When it has been so understood, the answers which have commanded most support are two.

First, there is the conventionalist answer which Dummett discusses in the article from which our quotation is taken, and the somewhat more general and looser 'linguistic theory' of necessity of which it is, perhaps, the sharpest formulation. According to this answer, necessary truths are truths guaranteed, somehow, by the meanings of the words we use to express them. Vixens cannot but be female because being female is simply part of what is meant by the word 'vixen.' And since the meanings of words are somehow a matter of convention – either explicit conventional stipulation or, perhaps more likely, tacit agreements put in place by speakers seeking to match their usage to that of their fellows – necessary truths may be held to be the product of such conventions. This kind of theory was at its most popular in the heyday of logical empiricism. It declined in popularity as the leading doctrines of that philosophy, and centrally, the notions of analytic truth and meaning, came under pressure, as well as being the target of direct attacks, first by Quine and subsequently by Dummett, which seemed to show that it must run into a vicious infinite regress. It clashes head on with the widespread acceptance, largely as a result of Kripke's lectures *Naming and Necessity*, of *a posteriori* metaphysical necessities such as 'Water is H_2O,' 'Gold is the element with atomic number 79,' and so on. But it continues to enjoy some support.[1]

Second, there are attempts to furnish a reductive explanation of modality in terms of possible worlds. Reduction requires that what constitutes a possible world be explained in wholly non-modal terms. Lewis's theory, discussed above, aims to meet this requirement

by taking a world *w* to be a spatio-temporally closed system of things – that is, anything spatio-temporally related to anything in *w* is itself in *w*, and nothing else is in *w*. The theory asserts that there is a huge number of such worlds – according to Lewis, at least $2^{2^{\aleph_0}}$ of them.[2] If the purported reduction is to have any plausibility, sheer numbers are clearly insufficient – the possible worlds postulated must differ from one another in ways which are plausibly sufficient to 'cover all the possibilities,' as we might put it. Lewis's theory seeks to achieve this by postulating worlds in accordance with a principle of recombination. So too does the alternative combinatorial theory, which purports to avoid the kind of extreme realism about worlds Lewis advocates, put forward by David Armstrong (see Armstrong, 1989). Very crudely, a *combinatorial* theory holds that the actual world is composed of some – a large number – of fundamental bits and pieces of some sort, and that any way of rearranging, or *recombining*, these bits and pieces constitutes a possible world. Lewis's and Armstrong's theories differ over what the fundamental bits are – for Lewis they are mereological atoms, or individuals, while for Armstrong they are 'thin particulars' and 'fundamental properties.' They agree in holding that any rearrangement of the fundamental bits gives us a world. Underpinning combinatorialism, for both Lewis and Armstrong, is an endorsement of a form of metaphysical atomism – a Humean denial that there are any 'necessary connections' between 'distinct existences.'[3] John Divers (2002) gives a clear and useful exposition and critical assessment of Lewis's theory (see also Divers and Melia, 2002; 2003).

Essence and Essentialist Theories of Modality

Although Dummett's first question *may* be taken as asking for a reductive account of modality, it is not clear that it *must* be. For one might think that even if there is no non-modal basis to which modal concepts, or modal facts, can be reduced, it may be possible to illuminate the source or ground of necessities and possibilities in general by reference to concepts and facts of a particular kind, even though those concepts and facts are not themselves explicable in non-modal terms. This is, perhaps, the best way to view what I shall call essentialist theories of modality.

An essentialist theory of modality should be distinguished from essentialism, as it is commonly understood – that is, as the doctrine, which goes back in one form to Aristotle, that things have some of their properties essentially, but others merely accidentally. Thus it may be held that Aristotle was essentially a man, but only accidentally a philosopher – he might well have never gone in for philosophy, but it is not possible that he should not have gone in for being a man. An essentialist theory of modality involves a commitment to essentialism, but goes beyond it, by seeking to explain necessity and possibility – or at least metaphysical necessity and possibility – in terms of essence: what is metaphysically necessary is what is true in virtue of the nature (or essence) of things, and what is metaphysically possible is what is not ruled out by the natures of things. Different such theories differ, in part, over how they interpret the *explanans* 'true in virtue of the nature (or essence) of' such-and-such. Broadly speaking, by a thing's nature or essence is meant what it is to be that thing. For example, it might held that to be a natural number is to be either zero or one of the successors of zero, or that to be water is to possess a certain physico-chemical structure, consisting in being composed of molecules in which two hydrogen atoms are bonded with a single oxygen atom.

Proponents of essentialism need not embrace an essentialist theory of modality. They may hold that what it is for a property to be essential to its bearers can be explained in modal terms. For example, the claim that whales are essentially mammals might be explained as equivalent to the *de re* modal claim that $\forall x \Box$(x is a tiger $\rightarrow$ x is a mammal).[4] According to an essentialist theory of modality, this gets things back to front – instead of trying to explain essence and essential properties by means of the usual modal operators, we should reverse the direction of explanation. Thus Kit Fine (see esp. Fine, 1994, but also Fine, 1995b; 1995a) agrees that there is an important connection between essence and necessity, but argues that attempts to define essence and essential properties in terms of *de re* necessity are bound to fail, because essential necessities are 'sensitive to source' in a way that *de re* necessities as such are not. For example, while it is necessary that 17 is a member of the set, {17}, whose sole member is 17, it is essential to {17} that it has 17 as a member, whereas it is not essential to 17 that it belongs to {17} – as Fine would put it, the necessity that 17 belongs to {17} has its source in the nature of the singleton set (or perhaps in the nature of sets in general), not in the number 17 itself. It is part of what it is to be {17} that it has 17 as a member, but it is not part of what it is to be 17 that it belongs to that set. Instead, Fine argues, we should take the notion of truth in virtue of something's nature or essence as basic. If we do so, we may define various kinds of necessity in terms of it. For any given kind of entity, there will be a class of propositions which hold true in virtue of the natures of entities of that kind – for example, we might identify mathematical necessities with those propositions which are true in virtue of the natures of mathematical entities.[5] And then, generalizing, we might identify metaphysical necessities with those propositions which are true in virtue of the natures of some entities or other. If, following Fine, we abbreviate 'it is true in virtue of the nature of x that p' to $\Box_x p$, we can express that it is metaphysically necessary that p by $\exists x \Box_x p$, or – allowing for the possibility that something may be true in virtue of the natures of several things – $\exists x_1, \ldots, \exists x_n \Box_{x_1, \ldots, x_n} p$.

Among the questions any essentialist theory must confront, two obvious and central ones are: Can the basic notion of a proposition's being true in virtue of something's essence or nature be made clear – or at least sufficiently clear to overcome the doubt and suspicion with which many philosophers have viewed it? And, can a theory of this kind meet the general epistemological challenge which, as we have seen, confronts any theory of modality? How is knowledge of essence possible? These questions, along with others they prompt, have been the focus of a significant body of recent work to which we can only allude here.[6]

Modal Knowledge

The epistemological challenge (in effect, Dummett's second question (see §1.4)) is one which any account of the nature and basis of modal facts must confront. It is, as Peacocke points out (Peacocke, 1999, pp. 3–4), a version of the same challenge – the Integration Challenge, as he labels it – as that posed by Benacerraf in the philosophy of mathematics – how to square a believable account of the nature of mathematical *truth* with a believable account of how it is *known* (see Benacerraf, 1973). Much recent work can be seen as attempting to meet it.

As Kant observed (Kant, 1781/1787, p. 43), experience may teach us that things are thus-and-so, but not what must be so. But it can, to a limited extent, tell us what is possible. In some cases, conclusions about what might or can be so may be inferred from entirely

non-modal premises – as the medievals put it, we may reason *ab esse ad posse*. From the fact that Dr Roger Bannister ran a mile in less than four minutes, it follows that it is possible that a human being should perform this feat. Obviously there are countless other conclusions about what is possible which may be inferred from premises about what is actual. But this is knowledge only of *realized* possibilities – the central problem of modal knowledge is to explain how we can have knowledge of *unrealized* ones, together with knowledge of *necessities*. It may be suggested that the *ab esse* route may give us at least some knowledge of unrealized possibilities from non-modal premises – that in some cases, we may argue that this table broke, from which it follows that it was possible that it should break, and putting this together with the additional premise that there are no relevant differences between this table and that other table (which has not broken), we may infer that that other table might break.[7] It is not, without closer investigation, clear that the requisite supplementary premise about absence of relevant difference will prove to be cleanly non-modal. But even if it can be argued to be so – perhaps it can be supported by evidence relating to many tables, some of which were observed to break and others not, and that our candidate table does not differ from those which actually broke in ways in which they do not differ among themselves, and does not differ from them in any of the ways the tables known to have broken do so – it is clear that this kind of inference gives us at best only severely limited knowledge of unrealized possibilities, namely possibilities of *kinds* which are realized. There is no prospect of its getting us to metaphysically interesting unrealized possibilities, such as the putative possibility that an individual might have had different origins, or have been composed of entirely different matter, or have been a frog rather than a man.

To get beyond this crippling limitation, some other approach is needed. Many, if not all, recent approaches can be divided into two broad groups – those which see our primary access to modal facts as consisting in recognition of possibilities, and those which see knowledge of necessities as primary.[8] Possibility-first approaches standardly see some form of conceivability as our guide to possibility. A clear example of a necessity-first approach is the simple inferential model of knowledge of necessity proposed in Kripke (1971).

The idea that conceivability or imaginability can give us knowledge of possibility has a distinguished pedigree.[9]A central question for this approach, on which Yablo's seminal paper (Yablo, 1993) focuses, is what we should take to be required for conceivability. A large part of this question concerns what constraints must be met if something is to be conceivable in the relevant sense. Most would agree that the putatively conceivable state of affairs must be specifiable without violating logical laws. But many would agree that this minimum condition is not sufficient. A plausible strengthening is that putative possibilities should be describable without doing violence to the non-logical concepts involved.[10] Taking this to be a sufficient as well as necessary condition for conceivability amounts to equating possibility with what is sometimes called broadly logical or conceptual possibility. Such an equation was made during the first half of the last century by logical empiricists and others who simply identified necessity with analytic truth (including Quine, see n. 4). Since there appears to be no conceptual incoherence in the supposition that gold is a compound or that water is an element, these – the contraries of familiar Kripkean examples of putatively *a posteriori* necessities – would qualify as genuine possibilities. If conceivability is to be seen as a route to knowledge of *metaphysical*, as distinct from merely conceptual, possibility, some further constraints are required.

Constraints which might be proposed include ones concerning a thing's substantial composition or constitution, its membership in some substantial kind, and perhaps its

biological or other origin.[11] The idea would be that facts about some or all of these matters should be held fixed in our attempts to conceive of counterfactual situations. Thus it may be held that if water is composed of H_2O molecules, then it is not properly conceivable that it should have a radically different chemical composition; that given that Aristotle was a man, we cannot properly conceive of him as having been a frog, and perhaps that given that Elizabeth II had her origin in a sperm and egg produced by George VI and Elizabeth Bowes-Lyon, no situation is conceivable in which she has a different biological origin. These constraints are all more or less controversial, the last especially so.

If we ask what might justify imposing such constraints, it is not easy to see what other answer might be given than that substantial composition or constitution, kind membership, and origin, are *essential* or *necessary* to things being the things they are. The fact, if it is one, that justifying the constraints involves appeal to modal facts does not directly show conceivability-based epistemology is not viable, for perhaps we can separate questions about what methods may be employed in acquiring modal knowledge from questions about their justification. But if that justification itself rests on underlying modal claims – about constitution, kind membership, origin, and perhaps other matters – it seems that at least some modal beliefs cannot be justified by facts about conceivability.

The general principles for which justification is required are of precisely the kind required for the development of the alternative, necessity-first, approach – at least if it is to proceed along the lines of Kripke's inferential model (see above §2.2) That is, they are general principles which imply the singular conditionals of the form $p \rightarrow \Box p$ which figure as the major premises for inferences to necessitated conclusions. According to Kripke, these major premises can be known *a priori*, 'by philosophical analysis.' A major task for this approach is to spell out just how, in convincing detail, that philosophical analysis (if that proves to be the best description of it) is supposed to go.[12]

The considerable recent literature in modal epistemology includes much critical discussion of the approaches briefly sketched here. In addition to works already cited, Gendler and Hawthorne (2002) is a useful collection of papers, and Vaidya (2011) provides a useful up-to-date survey.

Necessary and Contingent Existence, Actualism, and Possibilism

Almost everyone agrees that things might have been other than the way they actually are – that there are contingent facts. Many also think that the contingent facts include facts about existence – that at least some of the things which actually exist might not have existed, and that there might have existed things other than those which actually exist. This latter belief, about contingent existence, is known as *contingentism*. Opposed to it stands the view that everything which exists exists necessarily, and that nothing which does not exist might have existed – this is *necessitism*. According to contingentism, you and I, and countless other objects, might not have existed; according to necessitism, we and all other objects necessarily exist. Necessitism appears to fly in the face of common sense – surely you and I and many other things might quite easily not have existed? Necessitists reply by distinguishing between existing and being concrete. It is certainly true that we might not have been concrete embodied persons. But this does not mean that we might not have existed. What would we have been, had we not been concrete? The necessitist view is not that we would have been disembodied spirits. When we are not concrete, we are merely possible people – a

species of merely possible objects, or bare *possibilia*. Possibilism is the view that there are merely possible objects; actualists hold that everything is actual (i.e., that only what actually exists exists).

It will still seem to many that necessitism is an implausible doctrine. But some serious arguments have been advanced in its favor. It cannot simply be dismissed out of hand. One argument is from logical simplicity. It can be argued, plausibly, that the modal logic of absolute necessity should be S5, the simplest, but strongest, normal modal logic. In quantified S5, one can prove the Barcan principle and its converse, which together require that nothing that exists could have failed to exist, and that nothing which doesn't exist could have done so. To be sure, Barcan and Converse Barcan cannot be proved if the underlying logic is free. But necessitists may argue that simplicity considerations weigh against the complications that course entails. Another argument runs as follows: Necessarily, if I don't exist, the proposition that I don't exist is true. Necessarily, if the proposition that I don't exist is true, that proposition exists. But necessarily, if the proposition that I don't exist exists, then I exist. Hence necessarily, if I don't exist then I do exist. Hence, necessarily, I exist. These two arguments along with some others are given by Timothy Williamson.[13] The second has been met with various objections from defenders of contingent existents (see Rumfitt, 2003; Wiggins, 2003; Efird, 2010). Many philosophers find the possibilism necessitism requires unpalatable, and would probably view its implausibility as outweighing the simplicity arguments for necessitism.[14]

Notes

1 Perhaps the most concerted defense is Sidelle (1989), some of whose more recent work, Sidelle (2002), explicitly takes issue with post-Kripkean orthodoxy. Cameron (2010) seems also to advocate a form of conventionalism.

2 That is, at least as many as there are subsets of the continuum, or subsets of the set of all real numbers. See Lewis (1973, p. 90), where Lewis gives the number as at least $\beth_2$, which is the same as $2^{2^{\aleph_0}}$.

3 Cf. Lewis (1986, p. 87), Armstrong (1989, pp. xi–x). As interpreted by Lewis and Armstrong, and perhaps as intended by its author, Hume's denial of necessary connections between 'distinct existences' almost certainly forecloses against any serious form of essentialism (see Chapter 34, ESSENTIALISM, also below). Hale (2013, §3.3) takes issue with it on this ground. Direct reliance on Hume's denial might be avoided by taking a possible world to be any mathematically possible recombination of basic bits; but, setting aside the question whether an appeal to mathematical possibility compromises the reductive aim, it should be clear that the equation of *metaphysical* possibilities with mathematical ones (i.e., combinations which cannot be ruled out on purely mathematical grounds) is equally question-begging.

4 What makes this *de re* is that the necessity operator $\Box$ is applied to an *open* sentence, that is, a sentence containing a free variable. Contrast the *de dicto* modal sentence $\Box\forall x$ (x is a tiger $\rightarrow x$ is a mammal), in which the modal operator is applied to a *closed* sentence, with all occurrences of the variable x bound by the universal quantifier $\forall$. Some philosophers – most notably W. V. Quine (see, e.g., Quine, 1956) – have thought that *de re* modalities are unintelligible, and that we can, at best, make sense of *de dicto* claims because we can construe them as asserting that their embedded (closed) sentences are analytically true. According to this skeptical view, necessity resides in the way we talk about things, not in the things themselves, as Quine once put it.

5 Fine himself characterizes the mathematical necessities as the propositions true in virtue of the nature of the mathematical concepts and objects, and similarly for other kinds of necessity, such as logical necessities. The formulation in the text is better suited to the somewhat different essentialist theory put forward in Hale (2013, ch. 6).

6 In addition to works already cited, Wiggins (1976), Correia (2006; 2007), Shalkowski (1994; 1997), Lowe (2012), and Vaidya (2011) include discussions broadly in sympathy with an essentialist theory of modality or essentialism more generally. More critical discussion can be found in Mackie (2006), Noonan (2013), and Ahmed (2007). See also Della Rocca (1996), Gorman (2005), and Wildman (2013).

7 The example is taken from a paper by Sonia Roca-Royes (2017), who defends a version of this kind of modal empiricism, as she calls it.

8 For this classification, where the approaches are labeled possibility-first and necessity-first, see Hale (2002), which argues for the latter.

9 Most notably, in Descartes and Hume. For Hume, see especially Hume (1960). For Descartes, see, for example, Meditation VI: "First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it … I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and … a distinct idea of body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it" (Descartes, 1988, pp. 114–115). Descartes's position is complicated, and arguably rendered inconsistent, by his doctrine that God is the source or creator of all truths, including truths about what is necessary and possible. For some excellent discussion, focused on this problem, see Ian McFetridge's posthumously published, but sadly incomplete, essay "Descartes on modality" McFetridge (1990). A very useful commentary on Descartes's crucial letter to Mesland of May 1644 occurs in Anfray (2009, pp. 103–124).

10 Cf. Peacocke's modal extension principle (Peacocke, 1997, pp. 526–540; Peacocke, 1999, §4.2). Peacocke does not present his 'principle-based' account of modal knowledge explicitly in terms of conceivability; his idea rather is that our concept of possibility may be captured by a set of principles governing what counts as an admissible assignment of semantic values to concepts, and that our knowledge of possibility may be seen as the exercise of our grasp of these principles.

11 cf. Peacocke's constitutive principle for kinds (Peacocke, 1997, p. 540; or 1999, p. 145). Similar constraints play a key role in the interesting theory developed by Timothy Williamson, according to which knowledge of metaphysical modality can be obtained by essentially the same constrained imaginative procedure as he takes to give us knowledge of ordinary counterfactuals. See Williamson (2007, ch. 5, especially pp. 163–164).

12 Kripke himself supplies more detail in two cases – necessities of identity and (rather more controversially) necessities of origin. Some further attempts in the same direction are made in Hale (2013, ch. 11).

13 Williamson (2002; 1998; 1990; 2013). For a similar view, see Linsky and Zalta (1994; 1996).

14 An impressive defense of contingentism is Stalnaker (2012). Menzel (2014) provides a useful and quite comprehensive survey of the actualism and possibilism.

References

Ahmed, A. 2007. *Saul Kripke*. New York: Continuum.

Anfray, J.-P. 2009. *Qu'est-ce que la nécessité? Libraire Philosophique*. Sorbonne, Paris: J. Vrin.

Armstrong, D. 1989. *A Combinatorial Theory of Possibility*. Cambridge: Cambridge University Press.

Benacerraf, P. 1973. "Mathematical truth." *Journal of Philosophy*, 70(19): 661–680.

Cameron, R. 2010. "On the source of necessity." In *Modality: Metaphysics, Logic, and Epistemology,* edited by B. Hale and A. Hoffmann, pp. 137–152. Oxford: Oxford University Press.

Correia, F. 2006. "Generic essence, objectual essence, and modality." *Noûs* 40(4): 753–767.

Correia, F. 2007. "Finean essence and Priorean modality." *Dialectica*, 61(1): 63–84.

Della Rocca, M. 1996. "Recent work on essentialism: part 1." *Philosophical Books*, 37(1): 1–13.
Descartes, R. 1988. *Descartes Selected Philosophical Writings*, translated by J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.
Divers, J. 2002. *Possible Worlds*. London and New York: Routledge.
Divers, J., and J. Melia. 2002. "The analytic limit of genuine modal realism." *Mind*, 111(441): 15–36.
Divers, J., and J. Melia. 2003. "Genuine modal realism limited." *Mind*, 112(445): 83–86.
Efird, D. 2010. "Is Timothy Williamson a necessary existent?" In *Modality: Metaphysics, Logic, and Epistemology*, edited by B. Hale and A. Hoffmann, pp. 97–107. Oxford: Oxford University Press.
Fine, K. 1994. "Essence and modality." *Philosophical Perspectives*, 8: 1–16.
Fine, K. 1995a. "Ontological dependence." *Proceedings of the Aristotelian Society*, 95: 269–290.
Fine, K. 1995b. "Senses of essence." In *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*, edited by W. Sinnott-Armstrong, D. Raffman, and N. Asher, pp. 53–73. Cambridge: Cambridge University Press.
Gendler, T. S., and J. Hawthorne, eds. 2002. *Conceivability and Possibility*. Oxford: Clarendon Press.
Gorman, M. 2005. "The essential and the accidental." *Ratio*, 18(3): 276–289.
Hale, B. 2002. "Knowledge of possibility and of necessity." *Proceedings of the Aristotelian Society*, 103: 1–20.
Hale, B. 2013. *Necessary Beings: An Essay on Ontology, Modality, and the Relations Between Them*. Oxford: Oxford University Press.
Hume, D. 1960. *A Treatise of Human Nature*. Oxford: Clarendon Press.
Kant, I. 1781/1787. *Critique of Pure Reason*, translated by N. Kemp Smith. Basingstoke and New York: Palgrave Macmillan.
Kripke, S. 1971. "Identity and necessity." In *Identity and Individuation*, edited by M. K. Munitz, pp. 165–164. New York: New York University Press.
Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.
Linsky, B., and E. N. Zalta. 1994. "In defense of the simplest quantified modal logic." *Philosophical Perspectives*, 8: 431–458.
Linsky, B., and E. N. Zalta. 1996. "In defense of the contingently nonconcrete." *Philosophical Studies*, 84(2–3): 283–294.
Lowe, E. J. 2012. "What is the source of our knowledge of modal truths?" *Mind*, 121(484): 919–950.
Mackie, P. 2006. *How Things Might Have Been: Individuals, Kinds, and Essential Properties*. Oxford: Clarendon Press.
McFetridge, I. 1990. "Descartes on modality." In *Logical Necessity and Other Essays*, edited by J. Haldane and R. Scruton, pp. 155–212. London: Aristotelian Society.
Menzel, C. 2014. "Actualism." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/entries/actualism/(accessed August 27, 2016).
Noonan, H. 2013. *Kripke and Naming and Necessity*. London and New York: Routledge.
Peacocke, C. 1997. "Metaphysical necessity: understanding, truth and epistemology." *Mind*, 106(423): 521–574.
Peacocke, C. 1999. *Being Known*. Oxford: Clarendon Press.
Quine, W. V. O. 1956. "Three grades of modal involvement." In *The Ways of Paradox*. New York: Random House.
Roca-Royes, S. 2017. "Similarity and possibility: an epistemology of *de re* possibility for concrete entities." In *Modal Epistemology after Rationalism*, edited by B. Fischer and F. Leon. New York: Synthese Library.
Rumfitt, I. 2003. "Contingent existents." *Philosophy*, 78(306): 461– 481.
Shalkowski, S. 1994. "The ontological ground of alethic modality." *Philosophical Review*, 103(4): 669–688.
Shalkowski, S. 1997. "Essentialism and absolute necessity." *Acta Analytica*, 12(19): 41–56.
Sidelle, A. 1989. *Necessity, Essence, and Individuation*. Ithaca, New York: Cornell University Press.

Sidelle, A. 2002. "On the metaphysical contingency of laws of nature." In *Conceivability and Possibility*, edited by T. S. Gendler and J. Hawthorne, ch. 8, pp. 309–336. Oxford: Clarendon Press.

Stalnaker, R. 2012. *Mere Possibilities*. Princeton, NJ, and Oxford: Princeton University Press.

Vaidya, A. 2011. "The epistemology of modality." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/win2011/entries/modalityepistemology/ (accessed August 27, 2016).

Wiggins, D. 1976. "The *de re* 'must': a note on the logical form of essentialist claims." In *Truth and Meaning: Essays in Semantics*, edited by G. Evans and J. McDowell, pp. 285–312. Oxford: Clarendon Press.

Wiggins, D. 2003. "Existence and contingency: a note." *Philosophy*, 78(306): 483–494.

Wildman, N. 2013. "Modality, sparsity, and essence." *The Philosophical Quarterly*, 63(253): 760–782.

Williamson, T. 1990. "Necessary identity and necessary existence." In *Wittgenstein–Towards a Re-evaluation*, edited by R. Haller and J. Brandl, pp. 168–175. Proceedings of the 14th International Witttgenstein Symposium, vol. 1. Vienna: Holder-Pichler-Tempsky.

Williamson, T. 1998. "Bare possibilia." *Erkenntnis*, 48(2–3): 257–273.

Williamson, T. 2002. "Necessary existents." In *Logic, Thought and Language*, edited by A. O'Hear, Royal Institute of Philosophy Supplement 51, pp. 233–251. Cambridge: Cambridge University Press.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, T. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

Yablo, S. 1993. "Is conceivability a guide to possibility?" *Philosophy and Phenomenological Research*, 53(1): 2–42.

32

Relativism about Epistemic Modals

ANDY EGAN

1 Introduction

The paradigmatic instances of epistemic modals are the sorts of uses of "might," "must," and "possible" that occur in exchanges like the following:

> "Where are my keys?"
> "I'm not sure – they might be on the desk."
>
> "Where's Bob?"
> "He must be in his office. His light's on, and his jacket's hanging outside the door."
>
> "Does John have cancer?"
> "It's possible that John has cancer. He has some of the symptoms. They've run some tests, but we won't know the results until tomorrow."

These are uses of "might," "must," and "possible" whose truth is somehow or other bound up with something about somebody's epistemic state. Call these kinds of sentences, in which an epistemic use of "might," "must," or "possible" takes wide scope over an unmodalized clause, *simple epistemic modal sentences*, or *simple EM sentences*.

The overwhelmingly natural first thing to say about just how such sentences are bound up with our epistemic states – one that serves as the point of departure for most of the views about epistemic modals now in circulation in the literature – is that with these sorts of uses of "might," "must," and "possible," a simple epistemic modal sentence of the form *might:P* (or *possibly:P*) is true iff P is compatible with what's known, and *must:P* is true iff not-P is incompatible with what's known. The reason why epistemic modals are interesting, and why it's hard to give a satisfactory theory of them, is that it's remarkably difficult to transform that plausible-looking first shot into a worked-out account.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

The first obvious question to ask about such an account is, "known by whom"? And a brief survey of cases reveals that the answer has got to be, "it depends." A number of authors have presented a variety of cases that demonstrate this. (See, for example, Hacking, 1967; Teller, 1972; Kratzer, 1986; DeRose, 1991.) Here is a somewhat simplified version of one from DeRose (1991):

> *The Cancer Test*
> John's just had a test, a negative result on which would rule out cancer. Jane, not knowing what the results were, says (in response to an inquiry from Bill, who's heard a rumor that John has cancer), "It's possible that John has cancer – they've run a test that could rule it out, but we won't know the results until tomorrow." At the same time, the doctor, who has seen the negative results of the test, says, "it's not possible that John has cancer – we should start planning tests for other diseases."

It seems as if both Jane and the doctor speak truly. A plausible explanation of this is that it's Jane's knowledge that's relevant to the truth or falsity of Jane's claim, and it's the doctor's knowledge that's relevant to the truth or falsity of the doctor's claim. It's easy to construct similar examples for "might" (though it's complicated a bit by the inconvenient fact that there is no lexical item that clearly stands to "might" as "impossible" stands to "possible," which makes the examples a bit less clean).

The lesson that's typically been drawn from these kinds of cases is a *contextualist* one. Different utterances of epistemic modal sentences are responsive to different peoples' epistemic states because they are *about* different people's epistemic states. At a first pass: If Jane says, "John might have cancer," she speaks truly, because what she's said is *that it's compatible with what Jane knows that John has cancer*. If the doctor says, "John might have cancer," he speaks falsely, since what he says is *that it's compatible with what the doctor knows that John has cancer*. This is actually not quite the standard contextualist conclusion. The standard view has it that the relevant epistemic state needn't be the state of some *individual*, but could be (and often is) that of some group or community.

One way to motivate this is by drawing attention to a phenomenon observed by G. E. Moore (1962), in which he notes that it's possible to deny someone else's assertion of "it's possible that Hitler is dead by now" by saying, "I know he's not." It would be odd if an assertion about *my* knowledge could serve to deny your epistemic possibility claim if what epistemic possibility claims assert is just that the relevant proposition is compatible with *the speaker's* knowledge. It's to be expected, though, if what's being asserted is that it's compatible with the collective knowledge of a group to which we both belong. (There are complications about how to understand collective knowledge. See von Fintel and Gillies, 2011, for discussion.)

Another motivation comes from a family of cases discussed by Keith DeRose (1991), which are variations on *The Cancer Test* above. In one such case, Jane knows that the doctors have looked at the test results, but they haven't yet told her about them. Asked, "Is it possible that John has cancer?" she replies, "I don't know whether it's possible that John has cancer; only the doctors know. I'll find that out tomorrow when the results of the test are revealed." Here, it seems clear that the embedded possibility claim can't be about her knowledge – what she doesn't know is whether it's compatible with the doctors' knowledge that John has cancer, so the relevant group must include at least the doctors.

There is another family of examples that motivates variability in whose knowledge is relevant, largely inspired by (or anyway in the spirit of) a footnote in John Hawthorne's *Knowledge and Lotteries* (Hawthorne, 2006). These examples are meant to support the more radical hypothesis that a *single utterance* can be correctly evaluated as true by one assessor, and correctly evaluated as false by another, because the epistemic states that are relevant to correct evaluation are different for the two assessors. These kinds of examples figure prominently in the relativist arguments of, for example, MacFarlane (2011; 2014), Egan, Hawthorne, and Weatherson (2005), and Egan (2007).

Here is the relevant footnote:

> [A]s far as I can tell, ordinary people evaluate present tense claims of epistemic modality as true or false by testing the claim against their own perspective. So, for example suppose Angela doesn't know whether Bill is alive or dead. Angela says Bill might be dead. Cornelius knows Bill is alive. There is a tendency for Cornelius to say Angela is wrong. Yet, given Angela's perspective, wasn't it correct to say what she did? After all, when I say 'It might be that P and it might be that not P', knowing that Cornelius knows whether P, I do not naturally think that Cornelius knows that I said something false. There is a real puzzle here, I think, but this is not the place to pursue it further. (Hawthorne, 2006, p. 29, fn. 69)

And here is a case of the sort used in the relativist literature (this one is a modified version of one from MacFarlane, 2011):

> *Eavesdropping Brian*
> The evidence Sally and George have available to them leaves it open that Joe is either in Boston or in Berkeley. Brian, listening in from behind a nearby shrub, just saw Joe in Berkeley five minutes ago. Sally says to George, "Joe might be in Boston." George accepts Sally's assertion, and signals his assent by saying, "that's true." Brian rejects Sally's assertion, and signals his dissent by muttering to himself, "that's false."

George's assent and his attribution of truth to Sally's utterance seem completely in order. So do Brian's dissent and his attribution of falsity to the very same utterance. This suggests that it's not just that different epistemic states can be relevant to the evaluation of different utterances of epistemic modal sentences, but that different epistemic states can be relevant to different assessors' evaluations of the *same* utterance of an epistemic modal sentence. (As mentioned above, these cases are much more controversial than the cases meant to show variation in relevant epistemic state across utterances. We'll talk more about the objections and responses to these kinds of cases, and the arguments based on them, in §5.)

Pretty much all parties to the debate agree that we can't give a uniform, one-size-fits-all answer to the question of whose knowledge is relevant to assessing the truth or falsity of epistemic modal claims. Many also question whether it's really always *knowledge* that's at issue, or whether we should allow the relevant epistemic relation to vary as well. Here is a case from Hacking (1967):

> *The Salvage Ship*
> Imagine a salvage crew searching for a ship that sank a long time ago. The mate of the salvage ship works from an old log, makes some mistakes in his calculations, and concludes that the

> wreck may be in a certain bay. It is possible, he says, that the hulk is in these waters. No one knows anything to the contrary. But in fact, as it turns out later, it simply was not possible for the vessel to be in that bay; more careful examination of the log shows that the boat must have gone down at least thirty miles further south. The mate said something false when he said, "It is possible that we shall find the treasure here," but the falsehood did not arise from what anyone actually knew at the time. (Hacking, 1967, p. 148)

Cases such as Hacking's suggest that we sometimes want a relation *weaker* than knowledge, so that sometimes what's relevant to the truth or falsity of an epistemic modal claim is not what we *know*, but, for example, what we could come to know by some available method of inquiry. (Other cases – some of which are discussed in Dever (2013) – suggest that we sometimes want a relation *stronger* than knowledge, such that not everything that we know will always be relevant.)

In general, the truth or falsity of epistemic modal claims depends on the compatibility of the embedded proposition (the *prejacent*) with some body of information – the information that's within the epistemic reach of some person or group. Which person or group's epistemic reach is relevant, and what it takes for some piece of information to count as within a person's or group's epistemic reach, seems to be variable, varying at least with the context in which epistemic modal claims are made, and possibly also with the context in which they're assessed. This is why it's so difficult to turn the promising first shot at the truth-conditions of epistemic modal claims – that, for example, *might:P* is true iff P is compatible with what's known – into a worked-out theory.

One way to deal with this, and to allow for variation in which body of information is relevant, is to go contextualist, and to say that the relevant body of information is determined by the situation of the *speaker* – by the person who's uttering the sentence. Another is to go *relativist*, and say that the relevant body of information is determined by the situation of the *assessor* – by the person who's assessing the sentence for truth or falsity.

Our main focus here will be on relativism, and my goal here will be to put on clear display the outlines of the debate about relativism about epistemic modals. But it will be helpful to say a bit more about the structure of contextualist theories, since contextualism is the main competitor to relativism, and probably is (and ought to be) the default starting-point view. Accordingly, much of the motivation for relativism comes from the purported inadequacy of the contextualist options.

Two notes about things I won't do in what follows, before we move on:

First, there are (at least) two other options, which I won't discuss here for reasons of space. One option that's very much alive is to go with a theory according to which epistemic modal claims aren't in the truth and falsity business. (For advocacy of such proposals, see Yalcin, 2007; 2011; Swanson, 2011.) Another thing to notice is that both relativist theories and standard contextualist theories assume quite a close connection between semantic content and communicative import. We could go with a theory that separates these quite sharply, as do the sorts of *pluralist* theory offered (though not specifically about epistemic modals) by Cappelen and Lepore (2005), or the sort of propositional radical theory offered by Bach (2011).

Second, I'm going to follow the literature by focusing almost entirely on "might" and "must" – and much more heavily on "might" than "must." But it's worth bearing in mind Eric Swanson's (2011) warning that "might" and "must" are just two examples of a much broader category – the category Swanson calls "the language of subjective uncertainty" – which it would be nice to have a unified account of, and that there are theoretical risks associated

with focusing too hard on a limited range of examples. I will run these risks here in the interests of space and ease of presentation, but it's worth bearing them in mind as we move along.

In the next section, we'll look at some of the important features of contextualist views in general. We'll then look, in §3, at contextualist views about epistemic modals in particular. In §4, we'll discuss the internal workings of two different sorts of relativist theories. In §5, we'll discuss some standard arguments relativists have deployed to motivate relativism over contextualism, and we'll trace a few of the first steps of the ensuing dialectic of reply and objection. In that section, we'll look at a number of the standard replies to relativist arguments. At more or less every point in what follows, I will, unavoidably, leave out a lot. Rather than trying to lay out all of the fine details of the various branches of the dialectic, I'll be concerned primarily to lay out the central moving parts of the main positions in debate and to walk through the first few steps of the core arguments.

2 Contextualism

The following diagram (Figure 32.1) illustrates the more-or-less standard picture of how sentences and utterances come by their truth-values. (The canonical source of this picture is, of course, Kaplan, 1989. See also Lewis, 1980.)

A sentence is associated, in the first instance, with a *character*, which determines a function from contexts of utterance to contents expressed. (It's more standard to talk about the characters of lexical items than of sentences. But it's easy to assign a character to a sentence based on the characters of its constituent expressions.) A sentence that doesn't include any context-sensitive vocabulary will have a constant character – it will express the same content in every context of use.

Assigning truth-values to sentences in a context of utterance is a two-step process: first, the context of utterance serves as an input to the sentence's *character*, in order to determine a content. Then the *circumstance of evaluation* (or *index*) corresponding to the context of utterance serves as the input to the contextually determined content, in order to determine a truth-value. So the truth-value of a sentence in a context, or of a particular token utterance, is sensitive to the context of utterance twice over: first, as an input to character, and second, as the determinant of which index serves as the relevant input to content.

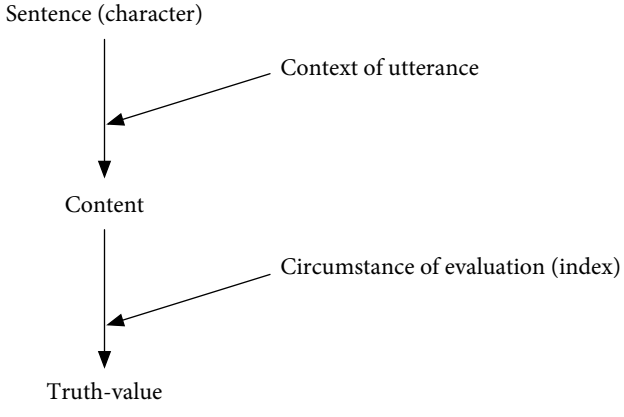

Figure 32.1

The key feature of all of this for our purposes is that standard contextualist views deliver a single, once-and-for-all verdict on the truth-value of a sentence in a particular context of use, and (therefore) on the truth-values of particular token utterances. While different utterances of the same sentence can take different truth-values, a particular, dated, world-bound utterance gets a truth-value *simpliciter*.

(I am going to assume – as I did above – that it's safe to move smoothly between issues about the assignment of truth-values to *sentences in context*, and issues about the assignment of truth-values to *particular token utterances*. I should note that this is potentially contentious.)

3 Contextualism about Epistemic Modals

There is a widely accepted, received view about the semantics of modals in general, the paradigmatic statement of which is found in Kratzer (1977; 1981). Here is, I think, a pretty uncontroversial, minimally formal way of stating it: The truth of *modal:P* at a world w depends on how things stand with respect to P in the worlds w* that bear some particular relation to w – the w*s *accessible* from w. (More generally, the truth of *modal:P* at an index i depends on how things are with respect to P at the indices accessible from i.) Different species of modality are distinguished by differences in the relevant accessibility relation. So what's nomologically possible in w is what's true in some world that's *nomologically* accessible from w – some world whose history is compatible with the laws of w. What's nomologically necessary in w is what's true in every world whose history is compatible with the laws of w. What's epistemically possible in w is what's true in some world that is, in some sense, *epistemically* accessible from w – plausibly, some world that's compatible with the epistemic state of some selected person or group in w. And so on.

(The above is not quite right: for a number of modals, we need not just an accessibility relation, but something that imposes an ordering on worlds. (See Kratzer, 1981; 1991.) We will ignore this complication in what follows.)

In this sort of framework, the distinction between, for example, metaphysical, nomological, deontic modals isn't a distinction between different sorts of lexical items, but between different kinds of uses of the same lexical items. There's just a single "might" in the lexicon, but it's context-dependent, since the relevant accessibility relation is different in different contexts. Similarly for "must," "possibly," and so on.

The overwhelmingly natural thing to say about *epistemic* modals, given this framework, and the thing we should say unless we're forced out of it, is that they are modals like any other, and context-dependent in just exactly the way that modals in general are context-dependent. Epistemic modals, that is, are just standard garden-variety modals, which happen to occur in a context that provides a distinctive sort of accessibility relation – one that has got something to do with somebody's epistemic state, such that the accessible worlds are the ones that some person or group stands in some particular epistemic relation to.

The simplest sort of contextualist account is probably what MacFarlane (2011; 2014) calls *solipsistic contextualism*, according to which the accessible worlds are those compatible with the speaker's knowledge. It's worth noting that this is a theory that makes decisions at two different choice points. It says that it's the *speaker*, rather than some other person or group, whose epistemic state is relevant. That's captured by calling it *solipsistic* contextualism. It also says that the epistemic relation that's important is *knowledge*, rather than some

other – possibly weaker, possibly stronger, possibly cross-cutting – epistemic relation. Different solipsistic contextualisms could differ on this point. So the view in question is probably better described as *solipsistic knowledge contextualism*.

The cases of variability we looked at in §1 suggest that solipsistic knowledge contextualism can't be right. DeRose's "might be possible" cases, and the facts about disagreement, suggest that the *solipsistic* part won't do, and cases like Hacking's suggest that the *knowledge* part won't do, either.

It should be clear, though, that solipsistic knowledge contextualism isn't the only option available for a contextualist about epistemic modals. There's a *lot* of room for variation in the story about just which people or groups, and which epistemic relations, are relevant, and in which circumstances. Pick any group, and any epistemic relation, that you like, and there's an accessibility relation that tracks what's compatible with what that group stands in that epistemic relation to. There is a lot to say about contextualism-internal debates about just which persons, groups, and epistemic relations are potentially relevant, and just how the relevance of a person, group, or epistemic relation depends on other features of the context. But what's important for our purposes is just the fact that contextualism offers quite a versatile framework for fleshing out the details of a theory of epistemic modals.

On our working contextualist theory, there will be two relevant contextually variable moving parts: the relevant group, and the relevant epistemic relation. An utterance of "Bob might be in his office" will always express a proposition of the type, *that it's compatible with all of the evidence that G stands in R to that Bob is in his office*, but it will express different such propositions in different contexts of utterance, because different Gs and Rs will be relevant in different contexts. The proposition that's expressed, that is, is a proposition about some person or group's epistemic state. It's true in worlds where the contextually relevant group fails to stand in the contextually relevant epistemic relation to evidence that rules out Bob's being in his office.

To sum up: this is a view according to which *epistemic* modals are just regular, garden-variety modals, used in a context that provides a certain distinctive kind of accessibility relation. What's distinctive about them is that their accessibility relations are ones that track what some person or group stands in some epistemic relation to. This allows for a lot of internal diversity within the category of epistemic modals, depending on whose epistemic reach the accessibility relation is tracking, and on what kinds of standards of *reach* are being applied.

4 Relativist Proposals

There are a number of different relativist views of epistemic modals on offer. All of them are committed to denying that utterances of simple epistemic modal sentences get truth-values "once and for all," and maintaining that a single, dated utterance can be true relative to one context of assessment and false relative to another.

There are two ways to do this. First, one can say that the context of assessment plays a role in the determination of the *content* of the utterance, so that a single utterance can express different propositions to different audience members. Call such views *content-relativist*. (See Predelli, 1996; 1998a; 1998b; Cappelen, 2008a; 2008b; and Egan, 2009, for defenses of such views – though not about epistemic modals.) Alternatively, one can say that the utterance has a single determinate content, fixed by its context of utterance, but that

(a) the content determines a truth-value relative to an index that includes not just a world, but also some further parameter with respect to which two contexts of assessment (within a possible world) might differ, and (b) which such index is relevant to assigning truth-values to an utterance can vary across contexts of assessment. Call such views *truth-relativist*.

I'm going to set content-relativist theories aside at this point, since the proposals about epistemic modals that have been discussed under the heading of "relativism" in the literature are predominantly truth-relativist proposals. So in what follows I'll use "relativist" to mean *truth-relativist*, and the relativist theories I'll be discussing here will all be instances of that type.

The various truth-relativist theories about epistemic modals all do two things, corresponding to (a) and (b) above: First, they postulate a further parameter, in addition to a world (and perhaps a time) in the index, or circumstance of evaluation, relative to which the contents of epistemic modal sentences take their truth-values. And second, they tell a story about the role this extra parameter plays in the notion of utterance truth, and in the practice of evaluating utterances for truth and falsity, such that we get the possibility of one assessor correctly evaluating an utterance as true, while another correctly assesses it is false.

John MacFarlane's relativism (MacFarlane, 2011; 2014) takes the propositions expressed by epistemic modal sentences to be functions from (at least) < world, information state > pairs to truth-values. (I'll write these '<w,q>' – using 'q' as a variable for information states and leaving 'i' to be used later on when we've got *individuals* appearing in our indices.) So the extra parameter is an information state – represented as a set of worlds. He then goes on to say that the process by which utterance truth is determined is more complicated than on the standard Kaplanian picture. Recall how we portrayed the Kaplanian picture before:

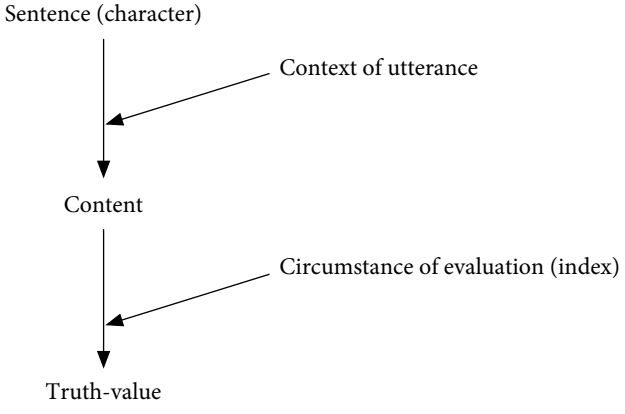

Figure 32.2

Really, this diagram leaves part of the story out. On the standard Kaplanian picture, not all indices are equal with respect to their role in the determination of truth-values for utterances, or for sentences in context. *Utterance* truth, and the accompanying notion of the truth of a sentence S in a context C, is sensitive only to the verdict that the content of S in C delivers about a *particular* index – the one that corresponds to the context C. (Depending on what parameters we've got in our indices, this is the index whose world is the world of C, whose speaker is the speaker of C, whose standards of precision are the standards in effect in C, etc.) If we want our diagram to reflect the full Kaplanian story about the determination of truth-values of sentences in context, we should write:

This addition allows our diagram to capture the way that, on the Kaplanian picture, the truth-value of a sentence in context is sensitive to context of utterance twice over – once as an input to character, and once as what we might call an *index-selector*. In particular, it indicates the special role that the context of utterance plays in the selection of the index that's relevant to the *sentence's* truth-value in the context of utterance, and so to the *utterance's* truth-value *simpliciter*.

Making this explicit allows us to clearly display the difference between a MacFarlanian relativist account and a standard Kaplanian account. On MacFarlane's sort of relativism, it's crucial that, when assessing an utterance of S in C for truth or falsity, the determination of which index gets fed in to the content of S in C in order to determine the truth-value of the particular utterance is sensitive, not just to the context of utterance, but also to the context of assessment. (If index-selection was only sensitive to the context of utterance, then we'd be unable to get any variation of utterance truth across assessors, no matter how complicated we made our indices. See MacFarlane, 2005; 2009.)

On MacFarlane's picture, which index is relevant to the determination of utterance truth is determined partly by the context of utterance, and partly by the context of assessment. For example, ignoring times and assuming we have just <w, q> indices, the relevant index for determining the truth of an utterance relative to an assessor is $<w_{CU}, q_{CA}>$ (where 'CU' names the context of utterance, and 'CA' the context of assessment).

The important difference here (see Figure 32.3), obviously, is that it's not the context of utterance alone that's responsible for selecting the index that's relevant to the assignment of truth-values to sentences in context, or to particular utterances. Index-selection, on this picture, is accomplished by the context of utterance and context of assessment together. As a result, this is a picture on which there's no such thing as the truth-value of a sentence in a context of utterance *simpliciter*, or of the truth-value of an utterance *simpliciter*. There's only the truth-value of a sentence relative to a *pair* of a context of utterance and a context of assessment, or of an utterance relative to a context of assessment. That's because it's only once we've got both a context of utterance and a context of assessment that we know which index to feed in to the content in order to get a truth-value. (On at least the first-pass version of MacFarlane's relativism about epistemic modals, *might:P* is true at <CU, CA> iff what is known to the assessor at CA is compatible with the truth of P at <CU, CA> .)

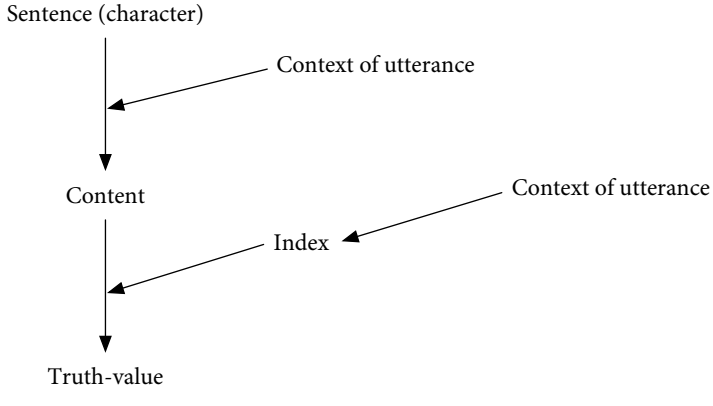

Figure 32.3

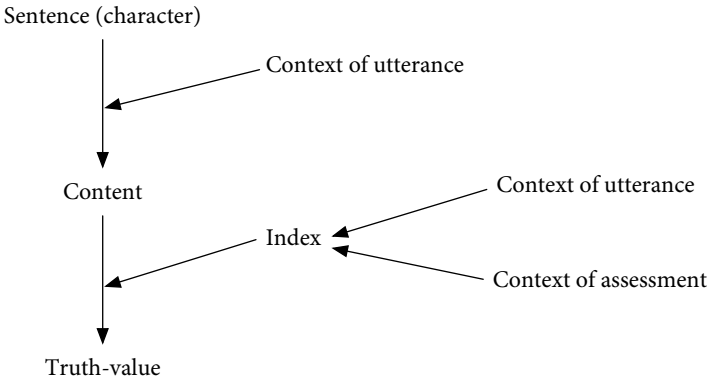

Figure 32.4

This allows for a different kind of variation in whose informational state is relevant than we find in contextualist theories. Here, the truth-values of epistemic modal claims are sensitive to the information available to the *assessor*. As a result, we can get the same utterance correctly assessed as true by one assessor, and correctly assessed as false by another.

Another relativist option is a theory that trades in *de se* linguistic content, as advocated by Tamina Stephenson (2007) and myself (Egan, Hawthorne, and Weatherson, 2005, Egan, 2007; 2010). Our views are different in detail, but for our purposes here I'll lump them together under the single heading of "*de se* relativism." On this kind of view, the parameter-adding is achieved by moving to a framework in which sentences in context have *de se* contents. (For more on the *de se*, see, for example, Lewis, 1979; Chisholm, 1981; Chierchia, 1989; Feit, 2008.) This is a picture of content on which the contents of sentences in context divide a space of possible *situations* or *predicaments*, rather than a space of worlds. Possible predicaments are standardly modeled with centered worlds, which we can think of as < world, time, individual > triples. (Peter Lasersohn's (2005; 2009) version of relativism is also in many ways similar, as is Max Kölbel's (2002; 2003; 2009), though their accounts are targeting other subject-matters. A third relativist option, which I will not discuss in detail for reasons of space, is Iris Einheuser's (2008) *factual relativism*.)

Possible predicaments (or centered worlds) differ from one another in many ways, including differing with respect to the evidential situation of the individual at the center. Epistemic modal sentences express centered-worlds propositions that are true of all and only those in a certain kind of epistemic situation. (For example, *might:P* is true at all and only those < world, time, individual > triples < w, t, i > such that the evidence within i's epistemic reach at t in w doesn't rule out P.)

Once we're given this account of the nature of the indices, we have the first component of a relativist story: we have contents that can take different truth-values relative to your predicament and to mine. We still need the second part of the story: an account of the role that the context of assessment plays in index-selection, such that it can happen that different indices are relevant to the assignment of a truth-value of an utterance when assessed by different individuals at different times. The simplest picture is this:

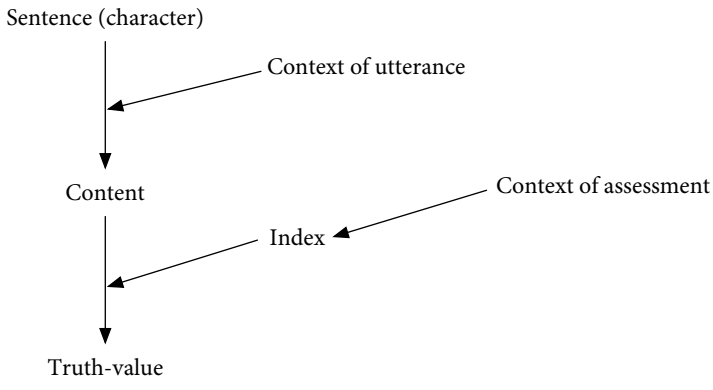

Figure 32.5

This is again a picture according to which the fundamental notions are not truth-in-a-context-of-utterance *simpliciter* for sentences, or of truth *simpliciter* for utterances. For sentences, the fundamental notion is truth relative to a < context of utterance, context of assessment > pair, where the context of utterance acts as input to character, and the context of assessment plays the index-selection role. For utterances, the fundamental notion is truth relative to a context of assessment.

This again makes the epistemic state of the assessor relevant to the truth and falsity of epistemic modal claims, in a way that it's not on the contextualist picture. And so it allows for variation in the truth-value of a single utterance relative to different assessors.

The above picture – on which the context of assessment always has sole responsibility for index-selection – is a natural one. But it's worth noting that it's not the only one available – this will be relevant later, when we're discussing some of the arguments for and against relativist theories.

One other thing that's worth noting briefly about the *de se* relativist view is that it can be stated in such a way that it's just an instance of the usual Kratzerian framework. The relativist needs to say that the relevant accessibility relations hold between centered worlds, rather than worlds, and they need to draw a distinction between "coarse grained" accessibility relations such that accessibility depends only on the *world* members of the < world, time, individual > triples (and so would be equally well modeled by accessibility relations between worlds), and those "finer grained" accessibility relations that are sensitive to other members of the triple. The proposal will then be that one of the distinctive features of *epistemic* uses of modal expressions is that the relevant accessibility relation is fine-grained, while other types of modality are governed by coarse-grained accessibility relations.

Notice that the kind of view just discussed is one that is both contextualist and relativist. The relevant accessibility relation is still fixed by context. But there's less diversity among epistemic accessibility relations contributed (at least in simple, unembedded environments) than standard contextualist views maintain.

5 Relativists' Arguments against Contextualism

Contextualist theories say that different utterances of the same sentence can take different truth-values, because they express different contents. Relativist theories say that there's more variation than this – that we can't accommodate all of the truth-value variation that we see in terms of variation, across contexts of use, in which contents are expressed.

There are two major families of arguments for this: First, what I will call *common-content* arguments, which aim to show that two utterances which we'd like to say differ in their truth-value must, after all, have the same content (and so that variation in truth-value can't be chalked up to variation in content). Second, what I'll call *single-utterance* arguments, which aim to show that particular, token utterances can take different truth-values relative to different assessors. We'll look at the common-content arguments first.

One family of argument for common content across occasions of utterance is based on *disquotational reporting*: If Fred says, "Bob might be in his office," it seems as if it's systematically okay for Fran to deliver disquotational indirect speech reports and (assuming Fred was being sincere) disquotational belief reports such as, "Fred said that Bob might be in his office" and "Fred believes that Bob might be in his office." This seems to be so regardless of what Fran's context is like – in particular, it's so regardless of whether Fred's and Fran's contexts are alike with respect to the features that will, given a contextualist account, plausibly be relevant to fixing the semantic value of "might." This isn't what you'd expect, at least at a first pass, given a contextualist account of "might."

Contextualist accounts seem to predict that "might" will take its semantic cues in the original utterance from *Fred's* context, and in the report from the *reporter's* context. If that were so, disquotational reports would run the same kind of risk of misreporting as do disquotational reports involving such paradigmatic indexicals as "I" and "here." If Fred says, "I am in New Brunswick," or "it's raining here," it *isn't* systematically okay for Fran to deliver disquotational reports. Reporting with "Fred said that I am in New Brunswick" is systematically bad, and "Fred believes that it's raining here" is only okay if Fred's and Fran's contexts of utterance are alike with respect to the semantic value they determine for "here."

In general, the problem is that Fred's utterance in C1 expresses the proposition *that Bob might$_{C1}$ be in his office*, while Fran's report in C2 expresses the proposition *that Fred said that Bob might$_{C2}$ be in his office*. Since C1 ≠ C2, it's very likely that "might" will express something different in the two contexts, and so it's very likely that Fran's report will be mistaken. (One natural possibility: Fran will misreport Fred as having said that it's compatible with what *Fran* knows that Bob is in his office.)

But disquotational reports are, in fact, completely safe. So, the relativist argument goes, there must not be a danger of "Bob might be in his office" having different semantic values in the context of utterance and the context of reporting. And so "Bob might be in his office" must have the same semantic value across the board. And so contextualism about "might" must be mistaken. And if contextualism about "might" isn't correct, then the best way to explain the variation that we see in the truth-values of different "might" claims is by saying that the stable content of "might" claims can take different truth-values relative to different contexts of assessment. (Versions of this argument occur in, for example, Kölbel, 2002, and Egan *et al.*, 2005.)

Another kind of disquotation-based argument for common content is based on disquotational attributions of agreement and disagreement. If Larry says, "Bob might be in his office," and Lisa says, "It's not the case that Bob might be in his office," Liz can report on this by saying, "Larry and Lisa disagree about whether Bob might be in his office." Similarly if Larry and Lisa both say, "Bob might be in his office," Liz can report on this by saying, "Larry and Lisa agree that Bob might be in his office."

These kinds of reports seem safe – their appropriateness doesn't seem to be hostage to the facts about what Larry's, Lisa's, or Liz's contexts are like. But we wouldn't – at least at first glance – *expect* these kinds of reports to be safe if contextualism about "might" were true.

For example, given a contextualist view of "might," there ought to be a danger that, since Larry was speaking in C1 and Lisa in C2, Larry said *that Bob might$_{C1}$ be in his office*, and Lisa said *that it's not the case that Bob might$_{C2}$ be in his office*. And these two claims, since they're (quite likely) about the epistemic situations of different persons or groups, needn't be in any conflict with each other. Worse, since "might" in Liz's report will get its semantic value from *Liz's* context, she's in danger of reporting Lisa and Larry as disagreeing about something that, quite likely, neither of them has any views about at all – for example, about whether it's compatible with what Liz knows that Bob is in his office.

Notice that the argument here isn't that, on a contextualist view, disquotational disagreement – or agreement – reports are *always* defective. The argument is that, in fact, the relevant kinds of disquotational reports are *always* okay, and contextualism predicts that they'll only be okay in certain narrowly constrained circumstances, in which the two parties' contexts are aligned with respect to the semantic value they determine for "might." Liz doesn't, in fact, have to verify that Lisa and Larry are in relevantly similar kinds of contexts before she can, with confidence, report them as disagreeing/agreeing. It's enough that she knows that they sincerely produced the relevant sentences. (Mark Richard (2004; 2009) makes these kinds of arguments in favor of a type of relativism in other domains, though he does not endorse relativism about epistemic modals.)

The arguments we've just canvassed are all arguments that the content of epistemic modal claims is stable across contexts of utterance, and so to the extent that we've got different intuitions about the truth-values of different utterances of a *might:P* claim, that must be because it's got the sort of content that can vary in truth-value at different points of evaluation within a world, rather than because it's expressing different contents in different contexts of use. Since the variation in truth-value across utterances isn't attributable to variation in *content*, it must be that the *same* content is taking different truth-values at the different indices relevant to assessment of the utterances.

This leads to a worry about all of these arguments: These look like better arguments for the sort of view MacFarlane (2009) describes as "nonindexical contextualism" than for any of the views that have been advocated under the heading of *relativism* about epistemic modals. They're arguments that different utterances of the same epistemic modal sentence (within a world) can take different truth-values, without differing in content. But they're not arguments that a *single utterance* of an epistemic modal sentence can take different truth-values relative to different assessors. And so they're not (at least not without further premises) arguments that the *assessor's* epistemic state ever needs to be relevant to determining the truth or falsity of an epistemic modal claim.

They're arguments, that is, that we need contents that take truth-values relative to something more than a world – since we've got the same contents being expressed by our various speakers in the examples, all of whom are worldmates, so that each of their contexts will supply the same world to the circumstance of evaluation that's relevant to the determination of utterance truth. But they're not arguments (at least, not without further premises) that particular utterances don't get their truth-values once and for all, or that we need to allow a role for contexts of assessment in the assignment of truth-values to sentences in context.

These sorts of common-content arguments are, in fact, most naturally understood as arguments that utterances with the same content can get different "once and for all" truth-values, *not* as arguments that we need to relativize utterance truth to contexts of assessment.

(For more on non-indexical contextualism, see MacFarlane, 2009. For extensive discussion and criticism of these sorts of arguments for the view that propositional truth is relative to more than a world, see Cappelen and Hawthorne, 2009.)

One argument, which is an elaboration of the disagreement argument above, *does* seem to support (if it's successful) not just stability of content, but also a role for context of assessment in the determination of truth-values for utterances. This is the argument from *faultless disagreement*.

There are two key moving parts of the argument from faultless disagreement: First, there's an imposition of a requirement that the relevant cases be classified by the correct theory as cases of *disagreement*. In the simplest arguments from faultless disagreement, this requirement is understood as a requirement that A's utterance semantically expresses the negation of B's utterance.

Second, there's a requirement of *faultlessness*. Both have to be getting it right, "from their own perspective." This isn't quite enough, however, as we've just seen that this by itself doesn't motivate a relativism of a kind that allows contexts of assessment to play a role in the determination of utterance truth. To accommodate this, all we need is a standard contextualist view (or a non-indexical contextualist view) that makes both A's and B's utterances come out to be true (true once-and-for-all).

But if we add a further requirement – we could call it a requirement of *strong faultless disagreement* – then we really do have an argument for a kind of relativism that allows utterance truth to be sensitive to context of assessment. To motivate this sort of relativism, we need to impose a requirement, not just that A is correct to regard his own utterance as true, but also that A is correct to regard B's utterance as *false*. (And *vice versa* for B.) This really does force us into a full-bloodedly relativist view, because then we have a pair of particular utterances – A's and B's – that need to take one truth-value for A, and the other for B (rather than just a pair of utterances that need to take different once-and-for-all truth-values, despite having the same content).

(This sort of argument features prominently in relativist arguments on other topics, but less in arguments for relativism about epistemic modals. For some applications, see Brogaard, 2008a; 2008b; Egan, 2010; Kölbel, 2002; 2009; Lasersohn, 2005; 2009; MacFarlane, 2007; 2014; Stephenson, 2007. For some opposition, see Sundell, 2009; von Fintel and Gillies, 2008; Dowell, 2012.)

This sort of argument from strong faultless disagreement is a member of a family of arguments meant to show that particular token utterances can be true relative to one assessor and false relative to another (within a world). That is, they're meant to show that utterances don't get truth-values once and for all, but only relative to a context of assessment. The argument from faultless disagreement bears a lot of the weight in discussions of relativism about other domains (in particular, about personal taste), but other members of the same family of arguments have standardly borne most of the weight in the case of epistemic modals.

One such argument (prominent in MacFarlane, 2011; 2014) is based on *retraction* phenomena. There's a common conversational phenomenon of criticism and correction, in which one party to the conversation takes another to task for some previous utterance. (I say, "The cat is on the mat." You say, "Nuh uh, the cat is not on the mat.") Abstracting away from a lot of detail, the standard responses to such criticism fall into two broad categories: *concession/retraction responses*, and *sticking to one's guns*.

Attempts at a certain sort of criticism and correction, which target utterances featuring context-sensitive vocabulary, after a relevant change of context has occurred in between the

original utterance and the challenge, often evoke a distinctive kind of sticking-to-one's-guns: an *impatience/insistence* response, marked by such replies as, "oh, come on," "you know that's not what I meant," and "don't be a jerk."

For example:

The Drive

Dave: "It's hilly here"
(They drive along, out into the plains, changing the contextual features relevant to determining semantic value of "here.")
Tim: "Nuh uh. It's not hilly here at all."
Dave: "Oh, come on. You know that's not what I meant. Don't be a jerk."

Note that the alternative concessive/retracting response from Dave sounds weird and inappropriate here – it would be bizarre for Dave to reply to Tim's challenge with any of the kinds of replies that mark this sort of response (such as, for example, "oh, I guess I was wrong," "I take it back," "my bad," or "okay, scratch that, then"). A similar pattern (I'll say a bit more about just what the pattern is in a minute) seems to hold for context-sensitive vocabulary in general. It's easy to construct parallel examples for "tall," "nearby," and so on.

What's happened here is that, between the utterance and the challenge, there's been a shift in the relevant features of the conversational context. Prior to the shift, at the time of the original utterance, the context was one in which assertions of "it's hilly here" would be true. After the shift, the context has changed in such a way that an assertion of "it's hilly here" made in the new, post-change context would be false. More generally: the context has changed in such a way that, while the speaker ought to have taken himself to be in a position to truly assert what he did in the original context, the speaker ought no longer take himself to be in a position to truly assert the same sentence in the new, post-change context. Let's call such a change in the conversational context an *undermining* context change, since such a change undermines the speaker's ability to felicitously re-assert the same sentence. (The above is my best effort to state the phenomenon in terms that are neutral between contextualism and relativism. I fear that I may not quite have succeeded, but I hope I have done well enough to allow you to identify the phenomenon.)

With epistemic modals, on the other hand, we see a different pattern of appropriate reactions to challenges after undermining context changes:

Jane Emerges

Jim: "Bob might be in his office."
(Jane, previously not a participant in the conversation, introduces herself into the conversation, bearing a photograph of Bob in Amsterdam, just posted to Facebook seconds ago, changing the contextual features relevant to fixing the semantic value of "might.")
Jane: "Nuh uh. There's no way Bob is in his office – see, here's a picture of him going into a bar in Amsterdam just seconds ago."
Jim "Oops, my bad – I take it back."

Note that the alternative, impatience/insistence response seems out of place here. It would be off for Jim to respond to Jane's challenge by saying, for example, "oh, come on – you know that's not what I meant."

Retraction arguments against contextualism (and in favor of relativism) about epistemic modals are based on this pattern of response to challenges and criticisms after undermining context changes, which don't look like the ones that a contextualist theory would predict.

The most interesting use of these examples, for our purposes, is as an argument that a single, dated utterance can get different truth-values relative to different contexts of assessment: Assessed from its original context of utterance, Jim's claim is true. Assessed from the post-shift context of assessment, the very same utterance is false. That's why the speaker *retracts*, rather than just refraining from uttering the same sentence in the new, changed context.

The challenge to the contextualist is to explain the retraction phenomena while insisting that the initial utterance retains the same truth-value relative both to the original context of utterance and the later context of retraction. The relativist argues that this is going to be impossible to do.

If the contextualist says that the original utterance was (and remains) *true*, it's difficult to explain why the speaker retracts in the face of the later challenge. If the contextualist says that the original utterance was (and remains) *false*, it looks as if she'll have to say that that's because Jane's information was relevant all along. And this, according to the relativist, is going to make the truth-conditions for "might" claims too demanding. Once we allow Jane to be (and to have been all along) a member of the group whose evidence makes a difference to the truth of Jim's claim, it's hard to identify a principled place to stop short of *everybody who ever thinks about the utterance* being in the relevant group. And if that's so, "might" claims are going to look incredibly risky, and we'll barely ever be in a position to felicitously assert them, since it will be so easy for them to be false.

Responses to this sort of argument from defenders of contextualism fall into two categories: challenges to the data, and attempts to accommodate the phenomenon within a contextualist theory. (Often, these moves complement each other – the contextualist offers a more sophisticated theory that accommodates a lot of the troublemaking phenomena, and then challenges the data to argue that the remaining un-accommodated phenomena aren't genuine, or aren't genuinely problematic.)

One compelling response of the data-challenging type is to point out that the tendency to retract isn't as universal as the initial versions of the relativist argument suggest. We don't *always* get retraction in the face of challenges after an undermining context change. We sometimes get insistence and impatience. Here is a case from von Fintel and Gillies (2008):

The Keys

Alex: The keys might be in the drawer.
Billy: (Looks in the drawer, agitated.) They're not. Why did you say that?
Alex: Look, I didn't say they *were* in the drawer. I said they *might* be there – and they might have been. Sheesh. (von Fintel and Gillies, 2008, p. 81)

Alex's response seems completely appropriate. He should, in this case, dig in his heels, refuse to retract, and insist on the correctness of his past "might" claim. So it looks as if the phenomena here aren't as the standard relativist arguments portray them – at least, they're not as the simple versions of the relativist arguments portray them. There are two ways to use this against the relativist.

One use of insistence phenomena is *defensive*: since the data aren't as the relativist described them, the arguments from retraction phenomena fail, and the world is still safe for contextualism after all. While it certainly does seem to be true that the data aren't as simple as the standard relativist arguments make them out to be, it's not clear how successful this reply is. Plausible versions of this response will still leave us with *some* cases of retraction after what looks like a relevant context change (in particular, after the introduction of new participants to the conversation). These retractions will still need explaining, and they aren't predicted by at least the first sorts of contextualist theories that spring to mind. So this can't be a *complete* defense – there's also a need for an elaboration on the contextualist theory that allows it to accommodate the remaining cases of genuine retraction. (One such theory is offered by von Fintel and Gillies, 2011.)

Another use of insistence phenomena is *offensive*: the insistence data are actually quite uncomfortable for the relativist. Relativist theories seem to predict that assessors' truth-value assessments will *always* be based on whether the utterance under evaluation is true relative to their own present context of assessment (or relative to the pair of the context of utterance and their own present context of assessment). They predict, therefore, that we'll *never* see the kinds of insistence on past correctness that we see in *The Keys*.

The relativist needs to either explain these away or provide a story that accounts for them. One way to offer such an account is to follow Peter Lasersohn's (2005; 2009) lead, and allow for two different sorts of truth-value attributions: one that tracks truth relative to the assessor's present context, and one that tracks truth relative to the speaker's context of utterance. In the cases of retraction, what's at issue is whether original utterance is true in the conversationalists' present contexts of assessment. Correctly recognizing that it's false relative to his own present context of assessment, the speaker retracts. In the cases of insistence, what's at issue is whether the original utterance was true in the original context of utterance. Correctly recognizing that it is, the speaker insists.

Relativists of both the *de se* and MacFarlanian types have the resources available to make this distinction. (Though in MacFarlane's framework it will be phrased slightly differently – as the distinction between truth at the pair $<C_U, C_A>$ and truth at the pair $<C_U, C_U>$ – and MacFarlane himself seems not inclined to adopt it.) Both also face the challenge of explaining, in a way that's not *ad hoc*, just why the one kind of evaluation is relevant in the retraction-generating conversations, and the other kind is relevant in the insistence-generating conversations.

Another sort of argument for variations in the truth-value of particular utterances is what are often known as *eavesdropper* arguments, based on the assessments of utterances by third parties who are not participants in the conversation in which the utterance takes place. Unlike what we find with standard context-sensitive expressions, and unlike what one would expect from a context-sensitive expression, evaluators who are not parties to the conversation in which an utterance occurs systematically base their assessments of utterances of epistemic modal claims on (their views about) their *own* evidential state, not on (their views about) the evidential state of the speaker, speaker's group, and so on.

For example, here is a contrasting pair of cases from MacFarlane (2011):

George and Sally

First case: You overhear George and Sally talking in the coffee line. Sally says, "I don't know anything that would rule out Joe's being in Boston right now" (or

perhaps, more colloquially, "For all I know, Joe's in Boston"). You think to yourself: *I* know that Joe isn't in Boston, because I just saw him an hour ago here in Berkeley. *Question:* Did Sally speak falsely?

Second case: Scene as before. Sally says, "Joe might be in Boston right now." You think to yourself: Joe can't be in Boston; I just saw him an hour ago here in Berkeley. *Question:* Did Sally speak falsely?

The hope is that you will have answered "no" to the first question and "yes" to the second. And this does indeed seem to be the reaction most people have. This example, as MacFarlane deploys it, is meant to do a number of things: It's supposed to identify an important difference between, for example, "Joe might be in Boston" and, for example, "As far as I know, Joe is in Boston," and thereby head off any equivalence claims that we might otherwise have been tempted to make. More importantly, it's also supposed to show that our assessments of the truth-values of other people's "might" claims, even those that take place in conversations that we're not parties to, are based on what we think about *our own* epistemic state, not on what we think about the epistemic state of the speaker, or the people who are party to the conversation in which the utterance takes place.

Here is another way of getting at the same phenomenon: Moore (1962) notes that one of the markers of the fact that "It's possible that Hitler is now dead" and "Hitler may be dead" are *epistemic* possibility claims is that one can use "I know that he's not" to deny them. This is an instance of a striking general feature of epistemic possibility claims, which is illustrated above, and also by Hawthorne's inspirational footnote: *anybody* can use first-person knowledge claims to deny them – not just addressees, and not just participants in the conversation in the course of which the claim is made.

The relativist argument here is that contextualists can't give an adequate explanation of this fact, while a relativist view is ideally suited to explain it. Contextualism, the relativist charges, predicts that competent extra-conversational assessors will defer, in their assessments of utterances for truth and falsity, to the context of utterance for content-determination and index-selection. So they shouldn't, in general, take their own epistemic state to be relevant, any more than they should take their own *location* to be relevant when they're assessing occurrences of "here" claims in conversations that they're not participants in.

One candidate contextualist reply here is to say the extra-conversational evaluator's epistemic state really *was* relevant all along, and so the utterance is just plain false. The trouble with this is that it's difficult to tell such a story without making the truth-conditions for epistemic "might" claims excessively demanding.

We don't, as a matter of fact, need to rule out the presence of better-informed lurkers in the closet in order to take ourselves to be in a good position to make "might" claims. And even if we *strongly suspect* lurkers – even if we're *certain* there's a lurker – we still don't take our "might" claims to be hostage to the lurker's knowledge. An example to make this clear: Tony Soprano can felicitously say to Paulie, "be careful, the Feds might have a guy outside your house," even if they're sure that there's an FBI agent listening in who knows for sure whether there's a guy outside Paulie's house. This is so even if Tony is pretty confident that there *isn't* a Fed outside Paulie's house, and that the eavesdropper knows this. And it's so even if he thinks the eavesdropper is hiding in the closet rather than listening in on a wiretap. (Tony could also say, "there might be a guy outside the house, and there might not. But it's best to be careful." At least one conjunct of this is sure to be false if the eavesdropper's knowledge is relevant.)

The relativist proposes to explain these phenomena by offering a semantic theory according to which a single, dated utterance of an epistemic modal claim can take different truth-values relative to different assessors, so that it's true for (e.g.) the speaker, and false for the extra-conversational assessor.

Here too, there are two types of reply available to the contextualist: they can challenge the data, or they can offer a way for a contextualist theory to accommodate the phenomena. (There is also the mixed strategy, of challenging the data to cut down the range of phenomena to be accommodated, and then telling a story that accommodates the remainder.)

Many of the standard presentations of eavesdropper arguments (see, for example, Egan, Hawthorne, and Weatherson, 2005; Egan, 2007) rely on intuitions about assessors' utterances of "that's true" or "that's false" in response to overheard epistemic modal claims. One concern about such intuitions is that just what they show depends on how some difficult issues about propositional anaphora wind up playing out. In particular, one way to resist the relevance of those intuitions is to offer alternative hypotheses about the reference of "that" in the relevant utterances, such that what the assessor is delivering an assessment of isn't the original utterance, or the proposition it expressed. (See, for example, von Fintel and Gillies, 2008, for some such replies.)

There are a couple of strategies available to the relativist in order to counter this reply. One is to tailor the responses in the examples in order to make the anaphora unambiguous (narrowing in, for example, on *Bob might be in his office* rather than the prejacent *Bob is in his office* as the target of assessment by noting that it seems in order for the evaluator to say, for example, "that's false, there's no way Bob is in his office," or "that's false, it's impossible that Bob is in his office"). Another alternative is to restate the argument in anaphora-free terms, as I have followed MacFarlane in doing here.

Besides challenging the relativist's claims about the data regarding extra-conversational assessments, another possible strategy is to adjust the contextualist theory in order to accommodate them. One option here is to tighten up the truth-conditions for epistemic "might" claims, so that the speakers *are* speaking falsely, but blamelessly. The intra-conversational assessors are making the same sort of blameless mistake as the speaker, and the extra-conversational assessors are just getting their truth-value assessments right. The most natural way to do this is to say that eavesdroppers are, in general, to be included in the relevant group whose epistemic state makes a difference to the truth or falsity of epistemic modal claims. This strategy is subject to the same kind of concern as the analogous response to retraction arguments above: it's very hard to do this without making the truth-conditions for epistemic modal sentences (a) so demanding that we're basically never in a position to assert them, and (b) demanding in ways that make bad predictions about when we should take ourselves to be in a position to assert them. (Though see von Fintel and Gillies, 2011, for a contextualist strategy that seeks to avoid this problem.)

There is, obviously, much more to say about all of this. But I hope that this section has succeeded in displaying at least the standard motivations for relativism about epistemic modals, and the first step or two of the ensuing debates.

Conclusion

I've tried in the preceding to lay out, in broad strokes, the candidate relativist views about epistemic modals, and the contextualist orthodoxy that they're seeking to replace. I've also tried to lay out, in similarly broad strokes, the main arguments relativists offer for

their views, and the first few steps of the ensuing debate between the relativists and the anti-relativists about whether or not those arguments succeed. While I've looked at some of the anti-relativists' responses to the relativists' positive arguments, one thing that I have not done, due to limitations of space, is survey the various positive arguments against relativism that its opponents offer. I will close, then, with some suggestions, for those who are interested, of points of entry into that area of the literature:

For some positive arguments against (or at least doubts and concerns about) relativism in general, and not specifically about epistemic modals, see, for example, Wright (2008), Cappelen and Hawthorne (2009), García-Carpintero (2008), Evans (1985), and von Fintel and Gillies (2008).

For some arguments against relativism about epistemic modals in particular, see, for example, Dietz (2008), Dowell (2010), von Fintel and Gillies (2011), Schaffer (2011), Yalcin (2007; 2011), and Swanson (2011).

References

Bach, K. 2011. "Perspectives on possibilities: contextualism, relativism, or what?" In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 19–59. Oxford: Oxford University Press.

Brogaard, B. 2008a. "Moral contextualism and moral relativism." *Philosophical Quarterly*, 58(232): 385–409.

Brogaard, B. 2008b. "In defense of a perspectival semantics for 'know.'" *Australasian Journal of Philosophy*, 86(3): 439–459.

Cappelen, H. 2008a. "Content relativism and semantic blindness." In *Relative Truth*, edited by M. García-Carpintero and M. Kölbel, pp. 265–286. Oxford: Oxford University Press.

Cappelen, H. 2008b. "The creative interpreter." *Philosophical Perspectives*, 22: 23–46.

Cappelen, H., and J. Hawthorne. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cappelen, H., and E. Lepore. 2005. *Insensitive Semantics*. Oxford: Wiley-Blackwell.

Chierchia, G. 1989. "Anaphora and attitudes *de se*." In *Semantics and Contextual Expression*, edited by R. Bartsch, J. van Benthem, and P. van Emde Boas, pp. 1–31. Dordrecht, Netherlands: Foris.

Chisholm, R. 1981. *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.

DeRose, K. 1991. "Epistemic possibilities." *Philosophical Review*, 100(4): 581–605.

Dever, J. 2013. "Epistemic modals." In *Routledge Companion to Epistemology*, edited by D. Pritchard, pp. 545–557. London: Routledge.

Dietz, R. 2008. "Epistemic modals and correct disagreement." In *Relative Truth*, edited by M. García-Carpintero and M. Kölbel, pp. 239–262. Oxford: Oxford University Press.

Dowell, J. 2010. "A flexible contextualist account of epistemic modals." *Philosophers' Imprint*, 11(14): 1–25.

Dowell, J. 2012. "Contextualist solutions to three puzzles about practical conditionals." In *Oxford Studies in Metaethics*, vol. 7, edited by R. Shafer-Landau, pp. 271–303. Oxford: Oxford University Press.

Egan, A. 2007. "Epistemic modals, relativism and assertion." *Philosophical Studies*, 133(1): 1–22.

Egan, A. 2009. "Billboards, bombs and shotgun weddings." *Synthese*, 166(2): 251–279.

Egan, A. 2010. "Disputing about taste." In *Disagreement*, edited by R. Feldman and F. Warfield, pp. 247–286. Oxford: Oxford University Press.

Egan, A., J. Hawthorne, and B. Weatherson. 2005. "Epistemic modals in context." In *Contextualism in Philosophy*, edited by G. Preyer and G. Peter, pp. 131–70. Oxford: Oxford University Press.

Einheuser, I. 2008. "Three forms of truth relativism." In *Relative Truth*, edited by M. García-Carpintero and M. Kölbel, pp. 186–203. Oxford: Oxford University Press.

Evans, G. 1985. "Does tense logic rest on a mistake?" In his *Collected Papers*, pp. 342–363. Oxford: Clarendon Press.

Feit, N. 2008. *Belief about the Self: A Defense of the Property Theory of Content*. Oxford: Oxford University Press.

von Fintel, K., and A. Gillies. 2008. "CIA leaks." *Philosophical Review*, 117(1): 77–98.

von Fintel, K, and A. Gillies. 2011. "'Might' made right." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 108–130. Oxford: Oxford University Press.

García-Carpintero, M. 2008. "Relativism, vagueness and what is said." In *Relative Truth*, by Manuel García-Carpintero and M. Kölbel, pp. 129–154. Oxford: Oxford University Press.

Hacking, I. 1967. "Possibility." *Philosophical Review*, 76(2): 143–168.

Hawthorne, J. 2006. *Knowledge and Lotteries*. Oxford: Oxford University Press.

Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Oxford University Press.

Kölbel, M. 2002. *Truth Without Objectivity*. London and New York: Routledge.

Kölbel, M. 2003. "Faultless disagreement." *Proceedings of the Aristotelian Society*, 104: 53–73.

Kölbel, M. 2009. "The evidence for relativism." *Synthese*, 166(2): 375–395.

Kratzer, A. 1977. "What 'must' and 'can' must and can mean." *Linguistics and Philosophy*, 1(3): 337–355.

Kratzer, A. 1981. "The notional category of modality." In *Words, Worlds, and Contexts: New Approaches to Word Semantics*, edited by H. J. Eikmeyer and H. Rieser, pp. 38–74. Berlin: De Gruyter. Reprinted in *Formal Semantics: The Essential Readings*, edited by P. Portner and B. H. Partee, pp. 289–323. Oxford: Blackwell, 2002.

Kratzer, A. 1986. "Conditionals." *Chicago Linguistics Society*, 22(2): 1–15.

Kratzer, A. 1991. "Modality." In *Semantics: An International Handbook of Contemporary Research*, edited by A. von Stechow and D. Wunderlich, pp. 639–650. Berlin: De Gruyter.

Lasersohn, P. 2005. "Context dependence, disagreement, and predicates of personal taste." *Linguistics and Philosophy*, 28(6): 643–686.

Lasersohn, P. 2009. "Relative truth, speaker commitment, and control of implicit arguments." *Synthese*, 166(2): 359–374.

Lewis, D. 1979. "Attitudes *de dicto* and *de se*." *The Philosophical Review*, 88(4): 513–543.

Lewis, D. 1980. "Index, context and content." In *Philosophy and Grammar*, edited by S. Kanger and S. Ohman, pp. 77–99. Dordrecht, Netherlands: Riedel. Reprinted in *Papers in Philosophical Logic*, pp. 21–44. Cambridge: Cambridge University Press.

MacFarlane, J. 2005. "Making sense of relative truth." *Proceedings of the Aristotelian Society*, 105: 321–339.

MacFarlane, J. 2007. "Relativism and disagreement." *Philosophical Studies*, 132(1): 17–31.

MacFarlane, J. 2009. "Non-indexical contextualism." *Synthese*, 166(2): 231–250.

MacFarlane, J. 2011. "Epistemic modals are assessment-sensitive." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 144–179. Oxford: Oxford University Press.

MacFarlane, J. 2014. *Assessment Sensitivity*. Oxford: Oxford University Press.

Moore, G. E. 1962. *Commonplace Book*. London: Allen & Unwin.

Predelli, S. 1996. "Never put off until tomorrow what you can do today." *Analysis*, 56(2): 85–91.

Predelli, S. 1998a. "I am not here now." *Analysis*, 58(2): 107–115.

Predelli, S. 1998b. "Utterance, interpretation and the logic of indexicals." *Mind & Language*, 13(3): 400–414.

Richard, M. 2004. "Contextualism and relativism." *Philosophical Studies*, 119(1–2): 214–242.

Richard, M. 2009. *When Truth Gives Out*. Oxford: Oxford University Press.

Schaffer, J. 2011. "Perspective in taste claims and epistemic modals." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 179–226. Oxford: Oxford University Press.

Stephenson, T. 2007. "Judge dependence, epistemic modals, and predicates of personal taste." *Linguistics and Philosophy*, 30(4): 487–525.

Sundell, T. 2009. "Conflict and Content." PhD diss., University of Michigan.

Swanson, E. 2011. "How not to theorize about the language of subjective uncertainty." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 249–269. Oxford: Oxford University Press.

Teller, P. 1972. "Epistemic possibility." *Philosophia*, 2(4): 303–320.

Wright, C. 2008. "Relativism about truth itself: haphazard thoughts about the very idea." In *Relative Truth*, edited by Manuel García-Carpintero and M. Kölbe, pp. 157–186. Oxford: Oxford University Press.

Yalcin, S. 2007. "Epistemic modals." *Mind*, 116(464): 983–1026

Yalcin, S. 2011. "Non-factualism about epistemic modality". In *Epistemic Modality*, edited by A. Egan and B. Weatherson, pp. 295–332. Oxford: Oxford University Press.

33

Internalism and Externalism

JUSSI HAUKIOJA

1 Introduction: Internal Duplicates and Supervenience

Suppose that somewhere – maybe on another planet, maybe in another possible world – an exact molecule-for-molecule replica of you exists, uttering the same sentences, in languages that are syntactically and phonologically identical to the ones you speak. Will the expressions uttered by your replica have the same meanings as the expressions you utter? Many people feel, at least pre-theoretically, that the answer is, obviously, "yes" – this is the semantic *internalist* answer. Yet, most philosophers of language today would say that the meanings of at least some expressions used by you and your replica may differ, and in so doing commit themselves to semantic *externalism*.

Semantic externalism and internalism can be characterized in various different ways. The central and distinctive externalist claim is, roughly, that the meanings of at least some linguistic expressions are not wholly determined by, or supervenient on, the intrinsic features of the speaker. In this chapter, I will understand internalism and externalism as supervenience theses, or rejections thereof. Semantic internalism is, then, the view that the propositional contents of the sentences uttered by a speaker supervene on the intrinsic features of the speaker. Semantic externalism, on the other hand, is the negation of semantic internalism: according to externalism, propositional content *fails* to supervene on the intrinsic features of the speaker.[1]

These characterizations are, of course, at most as precise as the expressions "propositional content" and "intrinsic features." Moreover, even when it has been made clear exactly how these expressions are here to be understood, the characterizations leave open a wide range of different views regarding which features of propositional content supervene, or fail to supervene, on the intrinsic features of a speaker, as well as which external features semantic externalism would claim have to be added to the speaker's intrinsic properties, in order to find a sufficient supervenience base for propositional content.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

One of my main aims in this chapter is to clarify what the main options for externalist theories are, and what sorts of arguments are relevant for a given kind of view.

Formulating the view explicitly in terms of supervenience helpfully illustrates the role of *internal duplicates* in arguments for externalism. Semantic externalism is typically *not* argued for by deriving it from other general principles or assumptions. Rather, the evidence typically cited as persuasive for externalism consists, at least primarily, of thought experiments featuring precisely the kinds of internal duplicates mentioned above. In such thought experiments – the most famous being, of course, Putnam's Twin Earth – we imagine internal duplicates that are situated in different environments, speaking languages that are syntactically and phonologically identical. The externalist appeals to an intuition, or judgment, according to which the semantic properties of the expression tokens produced by the two duplicates are *not* identical. If the externalist is right in this, some kind of semantic externalism follows: some semantic feature of at least some linguistic expressions is dependent on external features, and not supervenient on the intrinsic properties of the speaker.

Understood in this way, semantic externalism and internalism are claims in *foundational* rather than *descriptive* semantics.[2] That is, they are views regarding the determination basis of the meanings of linguistic expressions, not views concerning what the meanings of such expressions are. Unfortunately, in discussions of externalism and internalism this is not always carefully kept in mind, probably partly because the classic arguments for externalism were also arguments against semantic *descriptivism*, which in turn can be construed both as a thesis in foundational semantics *and* as a thesis in descriptive semantics. It is important, however, to note that the issues are distinct. For example, even though a descriptivist theory in descriptive semantics is typically combined with a semantic internalist view in foundational semantics, there is no logical inevitability here: one could consistently be a descriptivist about meaning but hold that the descriptive contents of our terms are partly externally determined (perhaps by committing to a social externalist view, cf. §4 below).

As noted above, semantic internalism can appear quite attractive, at least pre-theoretically. Indeed, internalism seems to have been the default position in philosophy of language until the 1970s. After the famous arguments by Kripke, Putnam, Burge, and others, however, the situation changed quite rapidly, and today most philosophers of language would subscribe to at least some kind of semantic externalist thesis. The main focus of the discussions has been on finding out exactly what kind of an externalist thesis (if any) the externalist arguments manage to justify. This is also very much visible in this chapter – my focus will be on different arguments for various kinds of externalist theses, rather than on arguments for internalism.

I have above characterized semantic externalism and internalism as general claims: internalism holds that propositional content is *always* supervenient on intrinsic properties – externalism, as the negation of internalism, holds that the corresponding expressions (i.e. expressions which are phonetically or orthographically identical) used by duplicates may *at least in some cases* differ in *some* of their semantic properties. But this general externalist claim is of course consistent with holding that the semantic properties of *some* expressions *are* supervenient on the intrinsic properties of the speaker. Thus, it is often useful to discuss externalist views about a *class* of expressions. One might, for example, be an externalist about natural kind terms (and thereby qualify as a semantic externalist in the general sense), but be an internalist about, say, artifact terms. I will use the terminology in both senses below: the context will, I trust, make the interpretation sufficiently clear.

I will proceed as follows. In the next section I will review the central thought experiments often considered as giving strong support to externalist theses, paying close attention to how internal duplicates figure in the experiments. In §3 I will distinguish between different semantic externalist claims, based on *which* semantic feature is being claimed to be at least partly dependent on features external to the speaker. In §4 different externalist claims are distinguished based on another, independent dimension: based on how the supervenience base for semantic properties should be widened, by taking in features of the external environment. Finally, in §5 I will look at methodological and meta-philosophical aspects of the internalism/externalism debate, and discuss what makes a particular kind of semantic externalist claim true, when it is true. Relatedly, I will look at whether and why thought experimentation should be thought of as a fruitful way of arguing for and against semantic externalism. For the most part, my aims in this chapter are clarificatory: to distinguish between different kinds of externalist claims and to clarify what kinds of argument are relevant in arguing for and against them. However, in my discussion of the methodological and meta-philosophical issues in §5 I will explore the consequences of adopting a *dispositionalist* and *meta-internalist* perspective, developed in recent work.

2 Origins of Semantic Externalism

The roots of semantic externalism are found in the works of Saul Kripke (especially Kripke, 1980) and Hilary Putnam (especially Putnam, 1975) – indeed, the label "Kripke–Putnam externalism" is sometimes used. Despite the many points of convergence, however, Kripke and Putnam developed their views independently of each other in the 1960s (for a recent account of the development, see Putnam, 2013). For contemporary externalism, the slightly later works by Tyler Burge (especially Burge, 1979) are just as central. In this section I will outline the central arguments for externalism by Kripke, Putnam, and Burge. I am going to be purposefully vague when it comes to different formulations of externalism – these will be clarified in some detail in §§3 and 4.

Unlike Putnam and Burge, Kripke does not in *Naming and Necessity* (Kripke, 1980) argue explicitly in terms of duplicates. His main aim is to refute *descriptivist* theories of meaning (or, rather, "naming") for proper names and natural kind terms; the role of internal versus external factors is not highlighted in Kripke's discussion. However, it is quite clear that the alternative causal-historical picture he sketches is externalist: the determination basis of meaning for these expressions involves external factors such as naming ceremonies and episodes of reference borrowing that may have taken place before the relevant speaker was even born. Moreover, in his discussion both of proper names and of natural kind terms Kripke argues that one might be in an *epistemic situation* that is *qualitatively identical* to the actual one, but situated in a different environment, and be referring to different entities or properties with one's proper names and natural kind terms (Kripke, 1980, pp. 102–105, 150–152). This is very much in the spirit of semantic externalism, formulated as a failure-of-supervenience thesis (although strictly speaking it leaves open the possibility that the semantic properties of proper names and natural kind terms are determined by intrinsic properties of speakers that are irrelevant to the qualitative aspects of the speakers' epistemic situations).

We can safely say that Putnam's Twin Earth thought experiment is the most famous argument for semantic externalism (although Putnam does not use the term "semantic externalism"

in "The meaning of 'meaning'"). Here, internal duplicates play an absolutely central role.[3] Putnam's main aim in the paper is to show that the following two traditional assumptions cannot both be true (Putnam, 1975, p. 219): (1) knowing the meaning of a term is just a matter of being in a certain (narrow) psychological state, and (2) the meaning of a term determines its extension.

To show that at least one of these assumptions has to be abandoned, Putnam presents his famous thought experiment (Putnam, 1975, pp. 223–224). Planet Twin Earth is very much like our Earth: we may even imagine that each of us has an internal duplicate on Twin Earth, sharing all our internal properties, our behavioral history, and so on. However, the liquid called "water" on Twin Earth does not consist of H_20, but of XYZ, where "XYZ" is an abbreviation for a complex chemical formula; XYZ is "indistinguishable from water at normal temperatures and pressures." Putnam goes on to imagine that a spaceship from Earth visits Twin Earth. At first the Earthlings will, according to Putnam, assume that "water" has the same meaning on Earth and on Twin Earth. When apprised of the chemistry, however, the Earthian spaceship will "report somewhat as follows: 'On Twin Earth the word "water" means *XYZ*'"[4] (Putnam, 1975, p. 223).

Next, Putnam imagines that in the year 1750, before chemistry had developed on either planet, a pair of intrinsic duplicates existed: Oscar on Earth and Twin Oscar on Twin Earth.[5] No speakers on either Earth or Twin Earth were aware of the molecular structures of the watery substances in their environment. Nonetheless,

> […] the extension of the term "water" was just as much H_2O on Earth in 1750 as in 1950; and the extension of "water" was just as much XYZ on Twin Earth in 1750 as in 1950. [Oscar and Twin Oscar] understood the term "water" differently in 1750 *although they were in the same psychological state* […] Thus the extension of the term "water" (and, in fact, its 'meaning' in the intuitive preanalytical usage of that term) is *not* a function of the psychological state of the speaker by itself. (Putnam, 1975, p. 224, emphasis in the original)

If Putnam is right about the extensions of tokens of "water" on Earth and Twin Earth, the failure of supervenience on internal features is evident – Twin Earth is tailor-made to establish that such supervenience fails.[6]

In the same paper, Putnam also introduces slightly different kinds of examples to illustrate what he calls the "division of linguistic labour" (Putnam, 1975, pp. 226–227). For example, he tells us that he cannot distinguish between elm trees and beech trees. Nonetheless, according to Putnam, "we still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of 'beech' in *both* of our idiolects." This is because Putnam, and many other speakers, are *deferring* to *experts* who belong to the same linguistic community. Again, Putnam claims that a Twin Earth-style experiment can be constructed: suppose that on Twin Earth, the extensions of "elm" and "beech" are switched. This difference entails that the *experts* on the two planets cannot be each other's internal duplicates, but since Putnam and Twin Putnam are both unable to tell elm trees and beech trees apart, there may be no internal difference between them. Again, if we agree with Putnam about the extensions of "elm" and "beech," as used by him and his twin, the failure of supervenience is apparent.

Putnam's focus in "The meaning of 'meaning,'" and other papers putting forward externalist views, was solely on language. It was quickly noted, however, that if Putnam's

arguments work for natural kind *terms*, then a similar argument looks overwhelmingly plausible for natural kind *concepts*, establishing a similar externalist conclusion about them, and the intentional mental states in which they figure (McGinn, 1977). Soon thereafter, Tyler Burge argued that a similar result holds far more widely than merely for names and natural kind terms and concepts.

In his most famous thought example, Burge (1979, pp. 77–79) introduces a subject – let us call him Bert – reporting to his doctor on his ailments. Bert believes (correctly) that he has arthritis in his wrists and fingers, and has a number of other true beliefs about his disease. In addition, however, he falsely believes that his arthritis has spread to his thigh – he mistakenly believes that arthritis can affect both joints and muscles, whereas in reality arthritis is specifically an inflammation of joints. Next, Burge has us imagine a counterfactual situation where a subject – let us call him Twin Bert – "proceeds from birth through the same course of physical events that [Bert] actually does, right to and including the time at which he reports his fear to his doctor" (p. 77). Twin Bert and Bert are internal duplicates, but in Twin Bert's linguistic community, the term "arthritis" has wider application: "physicians, lexicographers, and informed laymen apply 'arthritis' not only to arthritis but to various other rheumatoid ailments" (p. 78). Let us suppose that on this counterfactual usage, Bert's muscle ailment *would* be included in the extension of "arthritis." According to Burge, when Bert says, "I have arthritis in my thigh," he is saying something false (indeed, necessarily false), while Twin Bert, uttering the same sounds, is saying something true – despite the fact that the two speakers are internal duplicates. Again, if we agree with Burge on this, semantic externalism in the sense under discussion follows directly.

Burge's main focus was on contents of intentional mental states, on our attributions of contentful mental states, and on the phenomenon of incomplete understanding. However, his thought experiment is at least equally effective as an argument for semantic externalism. And, as Burge points out, his thought experiment does not depend on any features specific to the term "arthritis," or to names of diseases in general: instead of "arthritis," we "could have used an artifact term, an ordinary natural kind word, a color adjective, a social role term, a term for a historical style, an abstract noun, an action verb, a physical movement verb, or any of various other sorts of words" (Burge, 1979, p. 79). If Burge is right, a similar argument can establish some form of semantic externalism about most, maybe even all, linguistic expressions.

3 Which Semantic Feature Is Externally Determined?

A noteworthy fact about the arguments mentioned in the previous section is that they concern first and foremost *reference* and *extension*. The driving force behind the thought experiments is the conviction that syntactically and phonologically identical utterances made by internal duplicates can have different *truth-conditions*. Most people who are impressed by the Twin Earth thought experiment take the experiment (and others like it) to establish a claim about *meaning*: not only are the *extensions* of some linguistic expressions partly determined by features external to the speaker (in a sense to be made clear below), so is the cognitive content of said expressions, or, the contents that speakers communicate when using them. Let us call this view *externalism about meaning*. Of course, if we assume (with Putnam) that meaning determines extension, then an argument for the claim that extensions are externally determined will also directly be an argument for externalism about

meaning. In practice, things are more complicated, as we will see. Many theorists have thought it possible to accept Putnam's judgments concerning reference in the Twin Earth example, but try to accommodate them in a theory of meaning that is in some central respects internalist.

Nonetheless, to accept the Putnamean judgments (or 'intuitions') about Twin Earth is already to accept that *some* central semantic features do not supervene on the intrinsic properties of the speaker. We need, then, a formulation of the minimal view that is directly supported by Twin Earth-style thought experiments – a view that one is committed to merely by virtue of agreeing with Putnam that Oscar's tokens of "water" refer to H_2O, but not XYZ, both on Earth and on Twin Earth, while Twin Oscar's tokens of a phonetically identical term refer to XYZ, but not H_2O, on both planets. Let us call this view *externalism about extension*.

Externalism about extension is the claim that the referents or extensions (across possible worlds[7]) of corresponding expressions, used by internal duplicates, may differ. More explicitly, according to *in*ternalism about extension, if speakers A and B are internal duplicates, then their phonetically and syntactically identical expression tokens e_A and e_B have identical extensions in all possible worlds. Externalism about extension denies this.

While most philosophers (post-Putnam) are externalists in this sense, internalism about extension has also had its defenders. According to the *common concept strategy* (e.g., Mellor, 1977; Crane, 1991; Segal, 2000), Oscar's and Twin Oscar's tokens of "water" have the same meaning and the same extension across possible worlds, applying both to H_2O and to XYZ. The problem with the common concept strategy is that it appears to fly in the face of semantic facts: externalists take it to be simply *obvious* that the extension of Oscar's tokens of "water" includes H_2O and excludes XYZ, and vice versa for Twin Oscar. (Not everyone will agree that this is obvious. I will return to this question below, in §5.)

Some externalists about extension want to insist that meaning and content are nonetheless, in some respect, internal – or at least that there is a component of content, narrow content, that is supervenient on the intrinsic properties of speakers. We should keep in mind that Putnam already acknowledged that there are two "plausible routes" to take, as a reaction to the Twin Earth thought experiment (1975, p. 245) – one could reject either of the two traditional assumptions that his argument, if successful, shows to be incompatible. His own route – rejecting the assumption that understanding a term is a matter of being in a particular narrow psychological state – is not dictated by the thought experiment. One might, alternatively, hold on to this assumption and reject the assumption that meaning determines extension. Putnam (1975, p. 246) feels that his route better respects our established ways of talking about sameness and difference of meaning, but this seems to be partly a matter of terminological preference.

However, not all hinges simply on terminological choice: substantial claims concerning knowledge and understanding, of what is communicated, and so on, can also come into play. For example, proponents of causal descriptivism and rigidified descriptivism (e.g., Kroon, 1987; Lewis, 1997; Jackson, 1998a) claim that the cognitive content of "water" is identical between Oscar and Twin Oscar: "water" is short for something like "the (actual) watery stuff of our acquaintance" for both of them. Such a view would count as an externalist theory of extension but an internalist theory of meaning: the meaning of (say) a proper name or a natural kind term would be internally determined, in that the corresponding expressions used by internal duplicates would have identical meanings.[8] In effect, this is to say that the meanings of the expressions in question have an implicit indexical element,

which they inherit from the explicitly indexical expression "our acquaintance." Internal duplicates could then be said to use their expressions with the same meanings in the same sense as different speakers using the personal pronoun "I" are, in one sense, using it with the same meaning. When the relevant intensions are construed as functions from *centered* possible worlds to extensions (and truth-values), they will be shared between internal duplicates.[9]

The combination of an internalist and an externalist element about meaning and extension, respectively, is explicitly stated in two-factor theories such as Fodor's (1987) or Loar's (1988). More recently, two-dimensionalists such as Jackson (1998b) and Chalmers (2006) have distinguished between two different components of meaning, primary and secondary intensions, where the former are shared between internal duplicates, but the latter may differ (and do differ, in Twin Earth-style situations), due to differences in the physical and/or social environment. Again, this counts as externalism about extension, but if one holds (cf. Jackson, 2004) that what is communicated is determined by primary intensions, the view is internalist about (communicated) meaning.

As noted in the previous section, many philosophers follow McGinn and Burge in taking Twin Earth-style argumentation also to establish *cognitive* externalism, or externalism about *thought contents*. A similar distinction between externalism about the extensions of one's concepts, or about the truth-conditions of one's intentional mental states on the one hand, and externalism about the cognitive content of one's intentional mental states on the other hand can and should be drawn here, as well.[10]

4 How Should the Supervenience Base Be Extended?

Semantic externalism claims, then, that a speaker's intrinsic properties are not sufficient as a supervenience base for the semantic properties of the terms used by the speaker. Externalism, as a general failure-of-supervenience thesis, is silent on what should be added to intrinsic properties of a speaker, to find a sufficient supervenience base. At the same time, the classic thought experiments were already formulated in ways that strongly suggest particular views regarding what should be added to the supervenience base. The additions can be thought of either as relational properties of the speaker or as properties possessed by various objects, kinds, processes, and so on in the speaker's environment – for my purposes here this choice does not matter.

The Twin Earth thought experiment, if successful, establishes *natural kind* externalism, or *physical* externalism. This is the view that the semantic properties of natural kind terms such as "water" supervene on the intrinsic properties of speakers, *plus* some features of the natural world that the speaker is causally interacting with. Typically, such features are thought of as intrinsic properties of physical objects (such as molecular structures, DNA sequences, and so on). But there is no obvious reason why the relevant properties of the physical objects might not themselves be relational – indeed, in the case of biological species, this is how they should be thought of, given the current state of mainstream biology and philosophy of biology.[11]

There is an unfortunate tendency in discussions of natural kind externalism to assume that the relevant features *have* to be construed as micro-structures, or more generally as underlying intrinsic properties, and even to ridicule natural kind externalism for assuming an outdated and oversimplified view of the scientific facts. It is true that Kripke and Putnam

assumed, at least for the purpose of illustrating their externalist views, that the relevant external features are simple and unified underlying features. But as far as semantic externalism, construed as a failure-of-supervenience claim, is concerned, it simply does not matter whether the external features in the supervenience base are intrinsic properties of members of the kind such as micro-structures, or extrinsic ones such as lineages, or something wholly different. All that matters, for externalism as a general claim, is that the added features are external to the individual speaker.

Burge's "arthritis" argument and Putnam's "elm/beech" argument point to another kind of semantic externalist claim, *social externalism*. This is the view that the supervenience base for propositional content has to include other members of one's linguistic community. There are in fact two quite different versions of this claim. On the first, the extensions or meanings of a speaker's expressions are dependent on the intrinsic states of a definite set of other speakers – this is the case when there are clear *experts* that a non-expert speaker is deferring to. Both Putnam's elm/beech example and Burge's arthritis example fall into this category: Putnam is unable to tell elm trees and beech trees from each other, but manages to refer to elm trees with tokens of "elm" because his linguistic community includes experts (botanists, or maybe gardeners), who *are* able to distinguish the trees from each other.

On the second version of the social externalist claim, the extensions or meanings of a speaker's expressions are dependent on patterns of usage in the speech community of which one is a part, without an assumption that some individual speakers should be designated as experts, and deferred to as such. If Burge's argumentation can, as he claims, be extended to color terms and artifact terms, the social externalist view that follows would, arguably, fall in this category – no special expertise is needed, for example, to recognize sofas and distinguish them from non-sofas, or to distinguish the colors from each other.[12]

5 Why Should We Accept Externalism?

If externalism is true of at least some of the linguistic expressions we use, *why* is it true of them? Moreover, if different kinds of externalism (physical, social, and so on) are true of different expressions, what makes this the case? Why is it, for example, that physical externalism about extension (arguably) holds for "water," but not for "bachelor"?

Relatedly, what kinds of methods should we use in finding out which kinds of externalist theses (if any) are true, and of which expressions? We have seen above that thought experiments have in fact played a central role in the argumentation for externalist views, but is thought experimentation a good method? And, supposing that it is at least an acceptable method, is it the best method available to us, or can we improve our methodology by supplementing or even replacing thought experimentation with 'real' experimental methods, as recently urged by experimental philosophers?[13]

We saw earlier that thought experimentation concerning internal duplicates can at most establish externalism about extension, not externalism about meaning. Although the latter view follows from the former together with the widely shared assumption that meaning determines extension, it is possible to be an externalist about extension and yet deny externalism about meaning. But what about the former view, externalism about extension? It is practically certain that there are no actual pairs of speakers who are one another's internal duplicates, so we cannot empirically investigate such duplicates. And even if we could, how would we establish whether the extensions of their expression tokens were identical or not?

It is simply not clear how this should be done – the extensions of one's terms are not directly observable. Any concrete suggestion for how we *should* go about investigating whether the extensions are identical or not will already need to commit to substantial views in meta-semantics. Similarly for any attempt to judge whether, and why, thought experimentation should be thought of as a good source of evidence for and against externalism.

A common way to explain the role of thought experimentation in philosophical semantics is roughly as follows. In thought experiments we elicit *intuitions* about the reference of linguistic expressions in possible scenarios. If, when considering an imagined scenario, we have an intuition that a certain object, substance, or sample (etc.) would belong to the extension of a given term, then that intuition is evidence for theories which would include this object, substance, or sample (etc.) in the extension of the term, and evidence against theories which would not. Likewise, if intuition tells us that a given entity would *not* belong to the extension of a term, then that intuition will be evidence against theories which would include that entity in the extension of the term, and evidence for theories which would not.

But what, exactly, are such intuitions? And why should we think that intuitions – whatever they are – reliably inform us about the extensions of our terms? If we have learned anything from the meta-philosophical debates about intuitions during the last 10 years or so, then it is that the term "intuition" may be doing more harm than good here (Cappelen, 2012). In considering thought experiments such as Twin Earth, we are not (at least not obviously) struck by a mental state with a special phenomenology, compelling us to accept Putnam's verdicts about tokens of "water," as used by Oscar and Twin Oscar. Nor do (or should) we take the judgments we arrive at, concerning thought experiments, as infallible. Nonetheless, when considering such thought experiments, most philosophers do report arriving, with minimal effort, at judgments concerning the extensions of terms in the scenarios imagined. But the minimality of our effort does not tell us whether and why such judgments should be trusted as evidence.

The recent debates around experimental philosophy should make us careful in assuming that such judgments are shared by all, or even most, speakers in one's linguistic community. Whatever one's opinion is about the seriousness of the experimental challenges that have actually been presented so far, one cannot simply assume that one's own judgments are universally shared – we need to know what to conclude if it turns out that judgments about, say, the Twin Earth thought experiment are subject to variation between cultures, socio-economic backgrounds, gender, and so on, or that they are, for example, only shared by philosophers with a certain kind of education, or with certain theoretical preferences.

Various different reactions to such variation are possible (for a useful overview see Horvath, 2010), and it should not be assumed without argument that the same kind of reaction will be appropriate in all the subfields of philosophy where reactions to thought experiments have been experimentally studied. When it comes to philosophy of language, and the debate between internalism and externalism in particular, it will be helpful to take a step back and bring in yet another internalism/externalism distinction: the distinction between *meta*-internalism and *meta*-externalism about reference, introduced in Cohnitz and Haukioja (2013). These views are concerned with the question: What makes it the case that the semantic externalist and internalist views described in previous sections have the truth-values that they in fact have? Semantic externalism and internalism, as we have seen, are views about what kinds of factors are involved in determining the extension or meaning of an expression. Meta-internalism and meta-externalism are views about what *makes it the case that* such factors are so involved. Meta-internalism is the view that the fact that a

semantic internalist or externalist theory is true of a given expression is determined by some intrinsic properties of the speaker in question. Meta-externalism is the denial of this: according to meta-externalism, such facts are *not* determined by intrinsic properties of speakers.

This distinction, too, can be formulated in terms of duplicates. Are the same theories of reference and meaning always true of the phonetically and syntactically identical expression tokens used by internal duplicates? For example, is it possible that internalism about reference and extension is true of the tokens of an expression as used by A, while externalism about reference and extension is true of the corresponding tokens as used by A's duplicate, B? Or, less dramatically, is it possible that different externalist theories are true of the tokens used by A and B (maybe the different versions of social externalism mentioned in §4)? According to meta-externalism, this is possible, while according to meta-internalism it is not. A meta-internalist will claim that, although the *extensions* of syntactically and phonetically identical expressions used by A and B may differ (if semantic externalism is true), the extensions are determined in the same way, and thereby the same internalist or externalist theory is true of the expressions.

Semantic externalism and meta-externalism have generally not been carefully distinguished from each other. Accordingly, explicit statements of meta-externalism are not easy to find: however, some fairly clear commitments to meta-externalist views can be found, for example, in Cappelen and Winblad (1999) and Ludlow (2003). Moreover, an unstated meta-externalist background assumption seems to be present in some of the recent debates concerning the role of experimental data in philosophy of language.[14] Cohnitz and Haukioja (2013) argue that meta-externalism should be rejected, because it fails to explain the role that reference has in the explanation of successful communication: if meta-externalism were true, there might be systematic mismatches between what linguistic expressions refer to, and what information is successfully communicated by competent speakers using such expressions. I cannot in this chapter undertake a full critical evaluation of meta-internalism and meta-externalism. Rather, I will focus on the consequences that a meta-internalist view would have on the methodological questions raised above, and suggest that a certain kind of meta-internalist view can make good sense of the use of thought experiments in arguing for and against externalism about extension.

A *dispositionalist* version of meta-internalism claims that the fact whether a given semantic internalist or externalist view is true of an expression as used by a speaker is supervenient on that speaker's dispositions to apply and interpret that expression – including the speaker's dispositions to apply and interpret the relevant expressions in conditions that are non-actual, as well as the speaker's dispositions to revise his or her application and interpretation in response to empirical information.[15]

Let us look at examples. For some of the expressions we use, such as "bachelor," our patterns and dispositions of application and interpretation are unaffected by contingent features of the natural world around us. The properties that speakers – individually or collectively – associate with the expression are sufficient for determining whether the expression applies to a given individual or not. With other expressions we have dispositions to 'shift the burden' of determining their applicability partly to external factors. In the case of "water," for example, we have dispositions to evaluate the correctness of actual and counterfactual applications of the term according to whether or not the term is applied to samples which share the underlying structure of the substance that is causally connected in the appropriate way to our actual usage of "water" (or something similar: the details will

depend on one's preferred theory of reference). We also have dispositions to *re*-evaluate our application of such terms in the face of new empirical information about what the world is like.

To illustrate, suppose that we took a representative sample of bachelors, studied them empirically, and found out that every single one of them has a certain neural structure – call it N – while no individuals outside the sample have N. Would we then start to categorize people as bachelors in other possible worlds according to whether or not they have neural structure N or not? Surely not. That is just not how "bachelor" works: we would go on categorizing people as bachelors, in the actual world as well as in counterfactual ones, according to their age, gender, and marital status.[16]

According to dispositionalist meta-internalism, social externalism arises in a similar fashion, through deferential dispositions. For example, I can refer to elm trees with my term "elm," even though I cannot tell them apart from beech trees, on the basis of my dispositions to defer to people who can actually tell elms apart from other trees (e.g., botanists or gardeners). Should I find that my classification of trees into elms and beeches differs from that of an expert, I would immediately be disposed to revise my earlier application of the terms and align my usage with the expert. In other kinds of cases, I might not be disposed to defer to any particular speakers and their usage, but rather be disposed to align my usage with that of the majority, and further to revise my earlier applications, should I find out that my usage was not in line with the majority. The fine details of our dispositions to apply and interpret, and to revise our application and interpretation, may be controversial. But if dispositionalist meta-internalism is true, the true theory of reference for a given expression is determined by precisely such dispositions.

If such a dispositionalist meta-internalist picture can be developed in detail, it promises to make it quite non-mysterious why thought experiments should be a valuable source of data for philosophy of language. On this view, semantic internalism and externalism are ultimately claims about complex patterns of dispositions that we have with respect to the expression, or class of expressions, in question: patterns to apply and interpret them, as well as to re-evaluate our past application and interpretation. Twin Earth-style thought experimentation is precisely a project in which one introspectively accesses such dispositions. Such thought experiments can be understood as a kind of mental simulation: we imagine finding ourselves in various metaphysically possible situations and ask ourselves whether or not we would be disposed to apply a given expression to a given entity (or substance, state, process, etc.), and how, if at all, we would reconsider our classifications in the face of various imaginable empirical findings, or in the face of finding ourselves disagreeing with others in our community, and so on. In doing this, we are not relying on empirical evidence, but rather activating our own dispositions in simulation. Thought experimentation can give us direct, though not infallible, evidence about the kinds of dispositions we in fact have.

That the primary evidence from thought experiments concerns our linguistic dispositions is easily clouded by the fact we often *report* the results of thought experimentation as generalizations. As a response to the Twin Earth thought experiment, we say that our term "water" refers to H_2O but not to XYZ, and vice versa for Twin Oscar's tokens of "water." But in considering the thought experiment, we are primarily probing our own dispositions to apply and interpret the relevant terms in a range of particular situations, our dispositions to revise our usage in response to various possible empirical outcomes, and so on. I suspect that most people who have discussed the Twin Earth thought experiment with non-philosophers (or taught it to students) have found that, if speakers are at first

unsure what to say about the case, getting them to think about particular application and interpretation situations will help them to see the relevant patterns.[17] On the dispositionalist meta-internalist view I'm suggesting here, saying that our term "water" refers to H_2O but not to XYZ (given that our world is an H_2O world) is really to report a generalization over such patterns.[18]

None of this, of course, answers the worries raised above, about possible variation in the relevant kinds of linguistic dispositions – we cannot simply assume that everyone else in our linguistic community shares our dispositions. On the view I am suggesting, the facts about reference and extension are determined by the patterns of dispositions that speakers actually have, and the precise nature of such dispositions is, clearly, an empirical question. However, it is far from clear that the kinds of experimental setups that experimental philosophers have primarily been using – survey studies where non-philosophers react to thought experiments – are the best way to study such patterns empirically. The kinds of thought experiments that have figured prominently in the internalism/externalism debates, in particular, are *philosophers' tools*: experiments formulated by philosophers *for* other philosophers who are already familiar with the competing theories and who are thereby able to focus on the relevant features of the scenarios that are (often sketchily) presented. To get more informative data, more sophisticated experimental setups should be developed (cf. Cohnitz and Haukioja, 2015).

The above discussion has concentrated on our reasons to accept externalism about extension. As noted earlier, externalism about *meaning* does not get the same degree of immediate support from considerations concerning internal duplicates. Here, a lot will turn on exactly what we expect a theoretically useful notion of meaning to do. Evidence about language use and dispositions to language use, whether collected by thought experimentation or 'real' experimentation, will not be as directly relevant here as it is for internalism and externalism about extension: rival theories of meaning – for example, a Putnamean theory and a causal descriptivist theory – may agree completely on extensions across possible worlds, but disagree on which phenomena a theory of meaning should primarily explain, or how the relevant phenomena are best explained.[19]

* * *

When it comes to the debates between internalism and externalism about extension as well as meaning, the battle lines between the opposing views have been quite clear for some time. But when it comes to the foundational and meta-philosophical issues I have touched upon in §5, they are much less so. I have here suggested one way of thinking about the foundational and methodological issues – future progress in the first-order debates about extension and meaning will, I believe, crucially depend on increased understanding of such foundational questions, not least the proper role, if any, of experimental results of various kinds.

Notes

1 This particular formulation can be found, for example, in Jesper Kallestrup's excellent recent book on semantic externalism (Kallestrup, 2012, p. 62).

2 Descriptive semantics is concerned with what meanings, or semantic values, the expressions of a language have, and how the meanings of complex expressions are determined by the meanings of their parts. Foundational semantics is concerned with what makes it the case *that* a language has the descriptive semantics that it in fact has. (See Chapter 35, REFERENCE AND NECESSITY.)

Sometimes the label "semantic externalism" is also used of views within descriptive semantics, namely, the view that the semantic values of some expressions are object-dependent. For discussion of the relationships between the views, see Wikforss (2008).

3 Although Twin Earth is often presented as a planet in another possible world, in Putnam's original presentation Twin Earth exists in *our* world, "somewhere in the galaxy" (Putnam, 1975, p. 223). However, in the discussion leading up to his thought experiment, Putnam notes that according to the traditional view he wants to reject, the meanings of terms used by internal duplicates (speakers who are in "the same psychological state in the narrow sense") in different "logically possible worlds" should be identical (p. 221). In discussions of thought experiments like Twin Earth, it is common to assume that it does not make much difference whether we adopt a "cross-world" or an "intra-world" interpretation of the scenarios. The differences *can* sometimes be highly important, however, for example in evaluating versions of descriptivism that aim to capture Putnamean judgments about reference in such cases, but this issue will not be examined here. For simplicity, and since I'm understanding semantic internalism and externalism as claims concerning supervenience (and since supervenience is standardly understood as a modal notion), I will here mostly adopt the cross-world interpretation.

4 I find it almost certain that the Earthian spaceship would *not* give a report on the *meaning* of the term "water" on Twin Earth. Rather, they would be more likely to report along the following lines: "Contrary to what we first thought, there is no water on Twin Earth, but rather a remarkably similar substance, XYZ." I will return to this below, in §5.

5 In Putnam's original example, Oscar_1 and Oscar_2.

6 There are, of course, complications to be worked out, even if one agrees with Putnam. For example, there is the uncomfortable fact that we Earthlings consist of 50–65% water. This worry is typically quickly dismissed as merely having to do with Putnam's choice of example. While I agree that the worry can be side-stepped in this way, I do think that the ease with which we dismiss it points to a more fundamental worry concerning how we should understand internal duplicates in discussions of externalism: perhaps the respect in which internal duplicates should be *duplicates* should not be thought of as identity of physical constitution, but rather as something like *subjective indistinguishability* (cf. Farkas, 2003).

7 The qualification "across possible worlds" is crucial. Classical descriptivism is a paradigm example of semantic internalism, yet any descriptivist should be happy to grant that the *local* extensions of phonetically and syntactically identical *definite descriptions*, used by internal duplicates, may differ: accordingly, so will the extensions of the terms covered by a descriptivist theory.

8 Externalists about meaning do, of course, have arguments against such proposals (beginning with Kripke's semantic and epistemological arguments), but such arguments are no longer driven by thought experiments featuring internal duplicates. Accordingly, externalism about meaning goes beyond what is immediately established by Twin Earth-style thought experimentation.

9 A centered possible world consists of a possible world, an agent in that world (here, the speaker), and a point in time. For further discussion of details of a theory of this kind, see Kallestrup (2012, pp. 102–105).

10 Relatedly, John McDowell (1992) argues that the proper lesson to draw from Putnam's Twin Earth thought experiment is that psychological states should not be thought of as *narrow* states, but rather as partly individuated by external factors: on such a conception of mental states, one can accept both that meaning determines extension, and that grasping a meaning is a matter of being in a certain psychological state. Such cognitive externalism, just as Burge's, concerns the individuation of mental content, and is often called *content* externalism. A more radical kind of cognitive externalism, *vehicle* externalism, claims that the contentful mental states themselves extend "outside the skull" – Clark and Chalmers's "Extended mind" thesis (Clark and Chalmers, 1998; Clark, 2010) is arguably the best-known example of vehicle externalism. (The term "vehicle externalism" is due to Hurley, 1998.)

11 There is nothing very surprising about this – after all, Kripke argued for the essentiality of origin for material objects, including organisms, so in the case of proper names the relevant "underlying properties" will be relational ones.

12 An interesting recent extension of social externalism that cuts across this division is also worth noting. *Temporal* externalism claims that the contents of our expressions can be partly dependent on contingent future events, whether these be future decisions made by a group of experts, or future trends and patterns in the linguistic community as a whole. For a defense of temporal externalism, see Jackman (1999); for critical discussion see Brown (2000), Jackman (2005), and Stoneham (2003).

13 For an excellent critical review of experimental philosophy of language, see Hansen (2015).

14 See, for example, the exchange between Machery (2011) and Devitt (2011) – both seem to assume that the semantic facts are constitutively independent of the linguistic judgments and dispositions of competent speakers.

15 A full defense of dispositionalist meta-internalism would require a discussion of whether and how the Kripkensteinian arguments against dispositionalism (cf. Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING) are relevant. A lot will turn on how we understand Kripkenstein's challenge and his distinction between straight and skeptical solutions, on whether meta-internalism is a reductive view in the relevant sense, and related issues: this discussion clearly falls outside the scope of the present chapter. However, my suggestion for how the distinction between semantic internalism and externalism arises on a dispositionalist meta-internalist view is inspired by a dispositionalist response to Kripkenstein that makes use of dispositions on two levels: dispositions to apply and interpret linguistic expressions, as well as dispositions to re-evaluate and correct one's application and interpretation in the face of new empirical evidence about the world and/or about the linguistic behavior of other speakers (cf. Pettit, 1996; Haukioja, 2005).

16 Or, consider another version: suppose that we found that, say, 99.7% of unmarried adult males have N, while 0.3% do not; at the same time, we find N present in a tiny proportion of married females. Would we revise our categorization of this small minority of men as bachelors and instead include the married females? Surely not. (Thanks to Daniel Cohnitz for this twist on the example). Were we to make a similar discovery about the golden stuff, say that we find some of the metal we categorized as gold to have a different atomic number than 99.7% of the rest, while a small sample of a greenish looking metal turns out to have the same atomic number as the other 99.7%, we would, I think, revise our categorization of the 0.3% and consider the greenish stuff as gold.

17 Indeed, this is precisely what Putnam does with his story about astronauts reporting back to Earth. The crucial dispositional facts here concern the astronauts' retraction of their earlier applications of "water" to samples of XYZ, when they become aware of the relevant chemical facts (cf. n. 4).

18 The judgments we report are, then, generalizations over non-inferential responses. Here I disagree with Cappelen (2012), who insists that the crucial judgments we arrive at in response to philosophical thought experiments are ordinary inferential judgments. Cappelen bases this claim on the fact that philosophers typically, after having presented a judgment about an imagined scenario, go on to give *reasons* or *arguments* for the judgment. Putnam is no exception: after presenting the Twin Earth thought experiment he goes on to hypothesize that the difference in the extensions of Oscar's and Twin Oscar's tokens of "water" is due to the fact that the extension of "water" consists of samples that bear the same relation to most of the stuff we call "water." But I think this is to misconstrue the role of such hypotheses. Putnam is not presenting an argument, drawing on common ground, to *convince* his reader that our "water" refers to H_2O but not to XYZ; rather, he is presenting a *diagnosis*, that is, putting forward a hypothesis about which factors our usage of "water" is in fact sensitive to.

19 Empirical and experimental data will potentially be relevant here, too, but the details will heavily depend on theoretical considerations. For example, Nichols, Mallon, and Pinillos (2016) have recently argued, based on experimental evidence, that natural kind terms are systematically ambiguous between internalist and externalist readings.

References

Brown, J. 2000. "Against temporal externalism." *Analysis*, 60: 178–188.

Burge, T. 1979. "Individualism and the mental." *Midwest Studies in Philosophy*, 4(1): 73–122.

Cappelen, H. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Cappelen, H., and D. Winblad. 1999. "'Reference' externalized and the role of intuitions in semantic theory." *American Philosophical Quarterly*, 36(4): 337–350.

Chalmers, D. 2006. "The foundations of two-dimensional semantics." In *Two-Dimensional Semantics*, edited by M. García-Carpintero and J. Macià, pp. 55–140. Oxford: Oxford University Press.

Clark, A. 2010. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Clark, A., and D. Chalmers. 1998. "The extended mind." *Analysis*, 58(1): 7–19.

Cohnitz, D., and J. Haukioja. 2013. "Meta-externalism vs meta-internalism in the study of reference." *Australasian Journal of Philosophy*, 91(3): 475–500.

Cohnitz, D., and J. Haukioja. 2015. "Intuitions in philosophical semantics." *Erkenntnis*, 80(3): 617–641

Crane, T. 1991. "All the difference in the world." *Philosophical Quarterly*, 41(162): 1–25.

Devitt, M. 2011. "Whither experimental semantics?" *Theoria*, 72: 5–36.

Farkas, K. 2003. "What is externalism?" *Philosophical Studies*, 112(3): 187–208.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Hansen, N. 2015. "Experimental philosophy of language." *Oxford Handbooks Online*. DOI:10.1093/oxfordhb/9780199935314.013.53

Haukioja, J. 2005. "Hindriks on rule-following." *Philosophical Studies*, 126(2): 219–239.

Horvath, J. 2010. "How (not) to react to experimental philosophy." *Philosophical Psychology*, 23(4): 447–480.

Hurley, S. 1998. *Consciousness in Action*. Cambridge: Harvard University Press.

Jackman, H. 1999. "We live forwards but understand backwards: linguistic practice and future behavior." *Pacific Philosophical Quarterly*, 80(2): 157–177.

Jackman, H. 2005. "Temporal externalism, deference, and our ordinary linguistic practice." *Pacific Philosophical Quarterly*, 86(3): 365–380.

Jackson, F. 1998a. "Reference and description revisited." *Philosophical Perspectives*, 12: 201–218.

Jackson, F. 1998b. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson, F. 2004. "Why we need A-intensions." *Philosophical Studies*, 118(1–2): 257–277.

Kallestrup, J. 2012. *Semantic Externalism*. London: Routledge.

Kripke, S. A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kroon, F. 1987. "Causal descriptivism." *Australasian Journal of Philosophy*, 65(1): 1–17.

Lewis, D. 1997. "Naming the colours." *Australasian Journal of Philosophy*, 75: 325–342.

Loar, B. 1988. "Social content and psychological content." In *Thought and Content*, edited by R. Grimm and D. Merrill, pp. 99–110. Tucson: University of Arizona Press.

Ludlow, P. 2003. "Externalism, logical form, and linguistic intentions." In *Epistemology of Language*, edited by A. Barber, pp. 399–414. Oxford: Oxford University Press.

Machery, E. 2011. "Expertise and intuitions about reference." *Theoria*, 72: 37–54.

McDowell, J. 1992. "Putnam on mind and meaning." *Philosophical Topics*, 20(1): 35–48.
McGinn, C. 1977. "Charity, interpretation, and belief." *Journal of Philosophy*, 74(9): 521–535.
Mellor, H. 1977. "Natural kinds." *British Journal for the Philosophy of Science*, 28(4): 299–312.
Nichols, S., R. Mallon, and Á. Pinillos. 2016. "Ambiguous reference." *Mind*, 125(497): 145–175.
Pettit, P. 1996. *The Common Mind*. New York: Oxford University Press.
Putnam, H. 1975. "The meaning of 'meaning.'" In *Philosophical Papers*, vol. 2, *Mind, Language, and Reality*. Cambridge: Cambridge University Press.
Putnam, H. 2013. "The development of externalist semantics." *Theoria*, 79(3): 192–302.
Segal, G. 2000. *A Slim Book about Narrow Content*. Cambridge: MIT Press.
Stoneham, T. 2003. "Temporal externalism." *Philosophical Papers*, 32(1): 97–107.
Wikforss, Å. 2008. "Semantic externalism and psychological externalism." *Philosophy Compass*, 3(1): 158–181.

34

Essentialism

GRAEME FORBES

1 Concepts

The term "essentialism" in its popular usage is usually qualified in some way, as in "biological essentialism," "gender essentialism," "social essentialism," and so on. The three views just mentioned are typical: they are all views about human nature, and their general thrust is that in certain respects people *have* to be the way that they in fact are, in virtue of, respectively, their genes, their sex, or the social class to which they belong. Usually the respects in question are politically controversial though there are also interesting examples with no real political overtones, for instance, Chomsky's view that it is part of the "human essence" to be capable of learning only languages whose syntactic rules satisfy the constraints of certain "linguistic universals" (Chomsky, 1988, *passim*). The general idea here is that for each thing of a particular kind there are various apparent possibilities for it which are in fact closed off in virtue of its possessing such-and-such a property, where the property mentioned is characteristic of the kind of essentialism being propounded. For instance, a human being may be said to be unable to partake in interpersonal relationships of a particular emotional timbre because that human being is male.

Contemporary *metaphysical essentialism*, which is our main concern here, consists in a variety of more abstract doctrines of this broad sort. Certain apparent possibilities for things are argued to be not genuine possibilities for them. However, the possibilities are closed off not in virtue of features of the things concerned as specific to them as, say, social class is to human beings in contemporary Western societies, but rather, in virtue of the very nature or identity of the thing. To be more specific, we may explain the idea of a *metaphysically essential property* in terms of Aristotle's "essential/accidental" contrast. According to Aristotle, an accidental property is "something which may possibly either belong or not belong to any one and the self-same thing" (Aristotle, 1928, 102^{b}5ff.). This can be firmed up in two slightly different ways. Assuming that an accidental property of x is a property that x in fact possesses, we may say either

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

(1) *P* is an accidental property of *x* iff (i) *x* in fact possesses *P* but (ii) there is a way things could have gone according to which *x* lacks *P*;

or

(2) *P* is an accidental property of *x* iff (i) *x* in fact possesses *P* but (ii) there is a way things could have gone according to which *x* exists and lacks *P*.

We may then define *P* to be an *essential* property of *x* if *x* in fact possesses *P* but *P* is not an accidental property of *x*. Metaphysical essentialism is more fundamental than the specific kinds of essentialism mentioned above, since these latter typically depend upon (alleged) features of human nature which are themselves accidental, so in ways things could have gone in which human beings do not have those features, they would not have to have the "essential" properties that depend on them.

What hangs on the difference between (1) and (2)? The problem with (1) is that it threatens to make all properties accidental, for an uninteresting reason, at least in the case of things which might not have existed. As we will see, the examples of essential properties that have captured most attention in contemporary philosophy are properties which are *existence-presupposing*: necessarily, if *x* has one of these properties, then *x* exists. But if *x* is a contingent being, one which might not have existed, there are ways things could have gone according to which *x* does not exist, and consequently, in any of these ways things could have gone, *x* lacks all the candidate essential properties, thereby demonstrating that they are not essential to *x* after all. But this seems merely to miss the point, and so one turns to (2) as a superior definition. Unfortunately, (2) also has its flaws, chief among which is that it makes *existence* an essential property, since there is no way things could have gone according to which *x* exists and lacks existence. So neither (1) nor (2) exactly captures the notion we are after. Why this should be so is itself an interesting question to which we shall return. For the moment, we simply observe that if we are to stay with (2), we will want to distinguish between *trivially* and *non-trivially* essential properties, with existence as the paradigm of the trivially essential. How this distinction is to be drawn is itself non-trivial.

The other main concept with which we will be concerned is that of an *individual essence*. Intuitively, the essence of a thing is the collection of its features which determine its identity, which make it the specific thing it is rather than something else. One way of articulating this idea is embodied in the following definition:

(3) *e* is the (*individual*) *essence* of *x* iff *e* is a set of properties such that each member of *e* is an essential property of *x* and it is not possible for any other object *y* to possess all the properties in *e*.

However, (3) is subject to an irritating technical defect of the same flavor as those afflicting (1) and (2), for there is nothing in any of our definitions which warrants the phrase "*the* essence" in (3). That is, there is no apparent reason why there should not be two different sets of properties *e* and *e′*, each satisfying the condition for being an essence of *x*. So (3) really ought to begin "*e* is *an* essence of *x* iff…." We shall return later to work of Kit Fine which motivates a different approach to the ideas of essence and essential property by emphasizing these difficulties with (1), (2), and (3). But for the moment we shall stay with the modal approach, the definitions in terms of possibility and necessity.

2 Essentialist Theses and Arguments for Them

The contemporary debate about essentialism was provoked by writings of Kripke (1972; 1980) and Putnam (1975), with subsequent contributions by Fine (1977) and Wiggins (1980). According to Kripke, the origin of an organism *o* is essential to it, and the matter from which an artifact *a* is fashioned is essential to it; or at least, neither the origin of *o* nor the matter of *a* could be entirely different from what it actually is. According to Putnam, the fundamental physical properties of substances are essential to them: a particular substance could not have had a totally different fundamental nature. For instance, there is no way things could go in which an actual chemical compound comes into existence with a molecular structure quite different from the structure it actually has. According to Fine, it is essential to a set to be a set and to have the members it actually has: a set is not to be conceived of as a box, which actually has one range of members, the odd numbers, say, but, if things had gone differently, would have had a different range of members, the even numbers, say (Fine, 1981). Finally, Wiggins has argued for the view that the "natural kind" (if any) to which *x* belongs is essential to *x* (1980, ch. 4). Thus he, Wiggins, as a human being, could not have been a polar bear, or a forest, or a performance of Beethoven's Ninth Symphony.

These claims have struck an intuitive chord in many philosophers. If we grant the essential/accidental distinction in the first place, then it is plausible that it is essential to Wiggins that he is not a performance of Beethoven's Ninth, but not essential to him that he is a professional philosopher (he might have pursued a different career). However, the interesting question concerns not so much whether a particular essentialist thesis is true, but rather what principles are employed in drawing the essential/accidental distinction in particular cases. Indeed, an independently plausible account of the principles should feed back to help decide some of the more controversial theses.

For purposes of uncovering principles we can sort the various doctrines into two groups: those which posit essential connections between specific individuals, such as the connection between a set such as {0} and its member 0 and between a tree and the seed from which it originated, and those which make a thing's kind essential to it, such as the claims that sets are essentially sets and humans essentially human. We begin with the first group, involving essential connections between specific individuals, and we take as a stalking-horse Kripke's proposal about the essentiality of origin, as formulated in the following famous passage (1980, pp. 112–113):

> The question really should be … could the Queen – could this woman herself – have been born of different parents from the parents from whom she actually came? Could she, let's say, have been the daughter … of Mr and Mrs Truman?… We can imagine discovering [that the Queen was the daughter of Mr and Mrs Truman] … But let us suppose that such a discovery is not in fact the case. Let's suppose that the Queen really did come from these [her actual] parents… can we imagine a situation in which it would have happened that this very woman came out of Mr and Mrs Truman? They might have had a child resembling her in many properties … [perhaps] … even … a child who actually became the Queen of England and was even passed off as the child of other parents. This still would not be a situation in which *this very woman* whom we call "Elizabeth II" was the child of Mr and Mrs Truman, or so it seems to me. It would be a situation in which there was some other woman who had many of the properties that are in fact true of Elizabeth… How could a person originating from … a totally different sperm and egg, be *this very woman?*

The question is somewhat rhetorical, but as Kripke develops his example it has considerable force, and I have never seen a convincing counter-example to the underlying thesis. This thesis is that if an organism *o* originates from a cell *c* – a fertilized egg or *zygote*, in the case of human beings – then it is essential to *o* that it originate from *c*: *o* could not exist except by originating from *c*. We will call this thesis the *essentiality of origin*. However, to find the essentiality of origin plausible and alleged counter-examples to it unconvincing is one thing; but to have a theoretical explanation of why it is true is another.

To find such an explanation we can consider some of the consequences of rejecting the essentiality of origin. If the thesis is false, then even if in the actual world *o* develops from *c*, there is a way things could have gone, an *alternative possible world*, in the popular jargon, in which *o* exists, but as a result of developing from a different cell *c′*. Let *w* be such a world, and for the sake of the argument let us suppose that *o* is as similar as possible in *w* to the way it is in the actual world. Let us also grant that there are no special connections between how things can go for *c* and how things can go for *c′*, other than that *c* and *c′* cannot both give rise to the same organism in the same possible world. Then there is a world *u* where *c* and *c′* both give rise to organisms; we use *r* for the organism to which *c* gives rise in *u*, and *g* for the organism to which *c′* gives rise in *u*. Let us pick such a *u* with the special feature that *g* as it is in *u* is as similar as possible to *o* as it is in *w*. The total setup is given in Figure 34.1.

We have refrained from making any suppositions about which, if either, of the organisms in *u* is identical to *o*. But in the logic of the situation, there are only three possibilities: (a) *o* is identical to *g*; (b) *o* is identical to *r*; and (c) *o* is identical to neither *g* nor *r*. The consideration which favors the essentiality of origin is that all three of these options seem to have undesirable consequences, and so the hypothesis which generates them, that there are worlds where *o* develops from different cells (such as the actual world and *w* in the figure), is to that extent disconfirmed.

Postponing discussion of (a) for the moment, the problem with (b) and (c) is that the hypothetical non-identity of *o* and *g* is hard to accept, since *o* and *g* are *intrinsic* and *spatio-temporal* duplicates: by this I mean that they have the same nature and occupy the same places at the same times. Indeed, they need differ only in the extrinsic respect that *g* in *u* is existing in a world where *c* gives rise to an organism, while *o* in *w* is not, and in other extrinsic respects which are a consequence of that one. If we were asked to imagine a course of events which differs from *w* in that *c′* gives rise to an organism different from *o*, but that is the totality of the difference – no difference in any features of the two worlds is to be admitted, only a difference in the pure identity of the thing to which *c′* gives rise – we may wonder what content could be given to the idea that *c′* becomes *different* things in the two worlds. But if the idea that the organisms into which *c′* grows in the two worlds are different is of dubious intelligibility in this example, it is hard to see how throwing in *c* and an organism it develops into can make any difference to the intelligibility issue. The objection to world *w*, then, on either (b) or (c), is that it implies the existence of a numerical difference between entities in distinct worlds where there are no intrinsic features of these entities to support the posited difference.

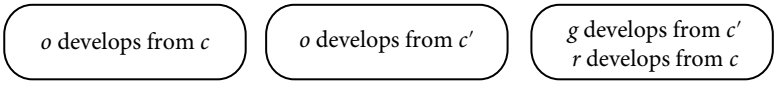

Figure 34.1

This still leaves the skeptic about the essentiality of origin with (a), that *o* is identical to *g*. However, under plausible assumptions, (a) has the same problematic aspect as (b) and (c). For there seems to be no reason why we cannot choose a world *v* in which *o* develops from *c* in the same way, and at the same places and times, as *r* does in *u*. That is, there is a world *v* in which *o* develops from *c* in such a way as to make it an intrinsic and spatio-temporal duplicate of *r* in *u*, since this only requires that the intrinsic and spatio-temporal features of *r* in *u* are ones all of which could have been possessed by *o*, excluding only those which involve the phenomenon at issue, such as being in a world in which something other than *o* develops from *c*.

This style of argument can be repeated for some of the other examples of essential properties we mentioned earlier. If we allow that a set could have different members, or that an artifact could have been made from entirely different parts, or that a substance could have had different fundamental physical properties, we can, under natural assumptions, generate analogous sorts of duplication without identity, as the reader may confirm. But this leaves us short of a *proof* that these properties are essential. Returning to the essentiality of origin, if intrinsic and spatio-temporal duplication of the sort manifested in our example implies identity, then any view which generates intrinsic and spatio-temporal duplication without identity is to be rejected. But this is insufficient to establish the essentiality of origin, since there are other properties which, if essential to *o*, would rule out intrinsic and spatio-temporal duplication without identity. For example, it might be proposed that the route through space which *o* traces while it exists is essential to *o*. If so, there is no world *v* such as is appealed to in the previous paragraph, and we can settle on (a). For in the actual world, in *w*, and in *u*, *o* must trace the same route; hence in *u*, *r* is spatio-temporally distinguishable from *o* in the actual world, and hence *r*'s route is not a possibility for *o*.

There is, of course, no plausibility in the hypothesis that an ordinary material object's spatio-temporal nature (its location at each time of its existence) is essential to it. The interesting question is why this is so. The kind of relationship in which an organism stands to its originating cell seems bound up with that organism's identity in a special way. It would go too far to say that the relationship is itself identity. A human being, for example, is not *identical* to the zygote from which he or she develops, since the zygote ceases to exist when it divides, while the human being does not cease to exist then. We may, if we wish, speak of a human's *zygotehood* on analogy with childhood (McGinn, 1976), but this simply raises the question of why the identity of the cell that constitutes the body during zygotehood should be any less accidental than the identities of the cells that constitute it during childhood. Nevertheless, the relationships between a set and its members, an organism and its zygote, and an artifact and the matter of which it is constituted (for example, a bronze statue and the bronze from which it is molded) do appear to have a certain affinity: they have an internal aspect which spatio-temporal relations lack. We will return to this later in our discussion of the source of necessity.

The kind of defense given here of the essentiality of origin is less obviously applicable in defense of Putnam's essentialism about the fundamental physical properties of substances. Could water exist in a universe where matter is continuous? Assuming that the fundamental property of water is not just to have a chemical composition of two parts hydrogen to one part oxygen, but rather to have molecules consisting in two hydrogen atoms and one oxygen atom, and hydrogen and oxygen themselves are fundamentally constituted of particles which are not further divisible, there could be nothing in a universe of continuous matter satisfying the description abbreviated by "H_2O." But perhaps water could exist there

as the substance fulfilling certain functions actually instantiated by H_2O. Suppose that we can make sense of the idea of functions that are performed by some substance in the actual world and again by some substance in a world where matter is continuous. Then we might argue, on Putnam's side, that this is still not sufficient for identifying the substances. But it will be hard to generate a convincing case of objectionable duplication that comes about as a result of making the identification, for this would require a world in which a substance with the physical structure H_2O exists alongside the substance from the world with continuous matter identified as water on functional grounds. At this point, one is less entitled to confidence in the coherence of the possible situations being stipulated than one is in those of the more straightforward case of organisms and the cells from which they develop. Still, this is only to say that a particular way of defending Putnam's thesis is less effective, not that the thesis itself is incorrect.

We turn briefly to the second group of essentialist theses we distinguished at the beginning of this section, those which say that a thing's kind is essential to it, such as the claims that sets are essentially sets and humans essentially human. These essentialist theses are easier to justify, particularly if we consider extreme cases. What could be meant by the suggestion that, say, {0} could have been a tree instead, or that Wiggins could have been Loch Ness? The problem of making sense of such bizarre hypotheses is not just one of lack of imagination. Indeed, there are fantasies in which persons in some sense "turn into" geographical features. But one does not treat these stories as representing genuine possibilities for the objects concerned; merely the same name is used, on the basis of some far-fetched or peripheral similarity. In the same vein, in explaining the layout of a town to someone over the breakfast table, one might say, "Let this pot of marmalade be the railway station." That does not mean that this pot of marmalade could have been a railway station, much less that particular railway station. The problem is that for it to be a possibility for a pot of marmalade to have been a railway station, or for {0} to have been a tree, we have to be able to conceive of two different states of affairs in which one and the same thing figures, in the first as a pot of marmalade or singleton zero, and in the second as a railway station or an item of flora, respectively. However, this in turn means that we would be conceiving of objects and their properties on the model of bare particular and inherence, according to which a thing is a propertyless substratum and can take on any nature you please via the inhering of appropriate properties. But this, if it makes sense at all, is at any rate not the conception which we employ. Articulating our actual conception is another problem, but whatever the right story is in this area (see Wiggins, 1980, chs 3–4), one constraint is that it must imply the fundamental unintelligibility of hypotheses which make the broad kind to which a thing belongs an accidental feature of it.

We have considered some examples of interesting essential properties and what might be said in support of them. An individual essence of an object x was defined in (3) to be a collection of essential properties of x such that if at any world an object y possesses all of them, then y is x. The essential properties we have discussed give rise to individual essences in a completely straightforward way only for sets: if x is a set, then *being a set* is part of x's 'natural' essence, and for each y which belongs to x, *having y as a member* is part of x's natural essence, and no other property is part of x's natural essence. By including *being a set* in the essence, we distinguish x from other entities which also have members and which might, as things actually are, have the same members as x, such as a club. The term 'natural' requires some justification, however. It is also an essence of x to be the sole member of $\{x\}$, at least by the lights of (3); but as Fine has urged (Fine, 1994) this essence is not revealing of x's nature: the kind of thing it is and *which* thing of that kind it is. So it is in that sense not the natural essence.

The analogous proposal for organisms is that each organism has its biological kind and the cell or cells from which it developed as its natural essence. But there are two difficulties. First, in view of the mechanisms by which speciation actually occurs, 'kind,' at least interpreted as 'species,' may be too strong. One would like something more vague here, but how can vagueness enter the specification of a thing's natural essence? Second, some cells arise by division of a parent cell. But if y and z arise in this way from x, all three being cells of some kind K, it is not sufficient to specify the essence of y as being of kind K and arising from x, for z is also of kind K and arises from x. However, y is constituted of part of the matter of x (just before division) and z is constituted of the other part, and it certainly seems wrong to say that y could have been constituted of the matter of which z is actually constituted, all else being the same so far as possible. Hence a plausible move in filling out the account of the natural essences of y and z is to add some constraint about what part of the matter of x each can be constituted of. But it also seems that it is too strong to make *exactly* the actual matter from which y is constituted part of y's natural essence. Surely y could have been constituted of slightly more, or slightly less, or slightly different, matter? Again there is the thought that we need to build some kind of vagueness into the specification of essences. We will see in the next section how this might be done.

3 Slippery Slopes and Primitive Thisnesses

The argument of the previous section claimed to ground certain essentialist theses, such as the essentiality of origin, in principles about "identity across possible worlds." The essentialist theses were defended on the grounds that denying them leads, under plausible assumptions, to pairs of worlds containing objects which are intrinsic and spatio-temporal duplicates and yet which are numerically distinct. But this is a poor defense if careful thought reveals that there is actually nothing objectionable about intrinsic and spatio-temporal duplication without numerical identity. We turn now to two arguments which seek to show that indeed there is nothing to object to in such duplication.

The first argument, due in essentials to Chisholm (1967) and sometimes known as Chisholm's paradox, exploits an intuition of *modal tolerance* in certain features of particular objects (the version I give of Chisholm's paradox is not quite Chisholm's, since he uses iterated modality; but the moral is the same). A watch, for example, which is actually made from a particular collection of parts, might have been made from a *slightly* different collection; it might have had a different winder, say. More abstractly, imagine that we have an artifact x made of parts $p_1, \ldots, p_{20}$, each of which is equally important to its function, and a different artifact y, though of the same design, made of parts $p_{21}, \ldots, p_{40}$. Suppose we now formulate modal tolerance more precisely as

(4) If a particular make-up is possible for a thing, then a slightly different make-up is also possible for it.

Since x is actually made of $p_1, \ldots, p_{20}$, then trivially, that make-up is possible for it (it could hardly be *impossible*!), and so by one application of (4), x could have been made of $p_1, \ldots, p_{19}, p_{40}$. But then by another application of (4) to this result, x could have been made of $p_1, \ldots, p_{18}, p_{39}, p_{40}$. By a further eight applications of (4), we arrive at the following conclusion:

(5) x could have been made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$.

However, exactly the same reasoning leads us to the conclusion that

(6) y could have been made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$

since y, being actually made of $p_{21}, \ldots, p_{40}$, could have been made of $p_1, p_{22}, \ldots, p_{40}$, and so on, using (4). Since there seems to be no reason why a situation in which (5) is true would have to differ in some other way from a situation in which (6) is true, other than in differences logically consequent upon the difference in identity of the things made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ in the respective situations, we have a striking case of duplication without numerical identity. But the logic that leads us to this point is impeccable, hence duplication without identity is unobjectionable.

One response to this is that in fact there is no duplication without identity, since any situation in which x is made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ is *ipso facto* one in which y is made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ and vice versa; it is just that in such a situation, x and y are the *same thing*. But this response is problematic in a number of ways. First, and most to the point in the present context, it is of little use to the essentialist, who will want to say that it is essential to x and to y to be made of a substantial majority of the parts of which they are actually made: a make-up of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ should be impossible for both x and y. Second, it has the strange implication that as things actually are, x is the same thing as any of the merely possible artifacts which could have been constructed using half of x's parts, for there are possible worlds with artifacts other than x with compositions that stand to $p_1, \ldots, p_{20}$ as x stands to the possible composition $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$. Third, on the usual way of understanding possible situations, it is hard to make sense of two things being identical in a possible situation. On the orthodox approach, we begin with a stock of objects whose possibilities we wish to model, and we represent possible situations involving them by selecting some of the objects and configuring them with atomic properties and relations; for instance, if x is selected and given the property P, and if y is also selected and x is given the relation R to y, the model is of the possibility for x of being P and being R to y. Thus if x and y are given from the outset as different things, then using x in the model of a possible situation is different from using y.

A better defense of the principle that intrinsic and spatio-temporal duplication implies identity is to query the logic of the supposed counter-example. It is natural to view (4) as a modal case of a paradox of vagueness, or a *sorites* paradox (see Chapter 28, SORITES). More familiar examples involve vague predicates like "tall" or "bald," or color words. For example, suppose 1,001 people are arranged in descending order of height, beginning with someone two meters tall and ending with someone one meter tall, adjacent people differing in height by one millimeter. The first person in this sequence is tall, the last is short. However, if we accept the seemingly plausible

(7) If x is tall then anyone not visibly distinguishable in height from x is tall

we easily derive, by repeated applications of (7), that the last person in the sequence is tall, though he or she is not. Case (4) is similar, the vague predicate being "_is a possible make-up for x." Whatever the problem is with relying on (7), then, the same difficulty should afflict (4).

However, it is one thing to know that a form of argument is untrustworthy, another to be able to say what theoretical flaw it embodies. There are various accounts of the mistake in the paradox of tallness, but a suggestive idea is that the argument depends on treating tallness as all or nothing: a person is either tall or not tall. Yet tallness is a matter of degree: someone can be tall, very tall, fairly tall, tallish, not exactly tall, and so on. Suppose we regard the various possible heights as correlated with *degrees of tallness* in such a way that different heights, even two that are not visibly discriminable, correspond to different degrees (unless they both qualify their bearers as tall *simpliciter* or not tall at all). The degree to which x is tall may be thought of as determining a *degree of truth* for the statement "x is tall," so that if we are somewhere in the upper third of our line-up, where the people are somewhat tall but also somewhat medium-sized, then "a is tall" will be true to a slightly higher degree than "b is tall" if a is one millimeter taller than b. Suppose also that the degree of truth of a conditional $p \rightarrow q$, when the antecedent has a higher degree of truth than the consequent, reflects the amount by which the truth of p exceeds the truth of q, the conditional's degree of truth dropping as the gap increases (in the limits, when p is not more true than q, $p \rightarrow q$ is wholly true, and when p is wholly true and q is wholly false, $p \rightarrow q$ is wholly false, as in classical logic; if we use the real numbers between 0 and 1 inclusive as degrees of truth, with 0 for complete falsehoods and 1 for absolute truths, then a clause for $\rightarrow$ which satisfies these conditions is: the degree of truth of $p \rightarrow q$, for short "$\deg(p \rightarrow q)$," is $1 - [\deg(p) - \deg(q)]$ when $\deg(p) > \deg(q)$, otherwise $\deg(p \rightarrow q) = 1$). The derivation of our conclusion that the one-meter-tall person in our line-up is tall can be represented as a chain of conditionals, "$\text{tall}(a_1) \rightarrow \text{tall}(a_2)$," "$\text{tall}(a_2) \rightarrow \text{tall}(a_3)$," and so on, the antecedent of the first conditional being given. However, there is now no reason to accept the conclusion "a_{1001} is tall," since somewhere in the first third of these conditionals we encounter ones in which the degree of truth of the antecedent is slightly higher than the degree of truth of the consequent. These premises are not wholly true, so the argument is unsound.

If this is a reasonable resolution of the paradox of tallness, how does it carry over to (4)? There are two different ways in which the transfer might be done, according to whether or not we wish to retain the standard method of modeling possible situations described earlier, which is to begin with a stock of objects whose possibilities we wish to model, and to construct possible situations by selecting some of the objects and configuring them with atomic properties and relations. If we retain this approach, another of its constituents which we can employ is that of the *relative possibility* relation between different possible situations. To use Nathan Salmon's terminology, each self-contained configuration of objects with properties and relations can be regarded as a *way* for those objects to *be* (Salmon, 1989: what follows is my own perspective on the approach of Salmon, 1986, which is not necessarily one with which he would agree). So we might have three possible situations, w, u, and v, where it is part of w that x and $p_1, \ldots, p_{20}$ stand in the "made of" relation, part of u that x and $p_1, \ldots, p_{11}, p_{32}, \ldots, p_{40}$ stand in the "made of" relation, and part of v that x and $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ stand in the "made of" relation. This gives us three different ways for x to be. But on top of this, we have to say which ways for x to be are possible relative to the various other possible situations, and this is a question which we can treat as a matter of degree. Relative to u, v has a high degree of possibility, indeed is perhaps *entirely* possible, since there is only a difference of one part in the make-ups of x in u and v. But relative to w, v can be stipulated to be not possible at all, while again relative to w, u may be regarded as a very remote possibility, though slightly more possible than v, because there is less difference in x. A possibility statement is read as asserting possibility relative to the actual world, so looking at (4) in this

light, we construe it as saying that if a situation in which x has such-and-such a make-up is possible relative to the actual world, so is a situation in which x has a slightly different make-up. But analogously to the case of people of different heights, it is possible to choose two make-ups for x such that x's having the second makes any situation which includes that as a way for x to be more remote from actuality than a situation which includes x's having the first needs to be. Repeated applications of (4), then, essentially involve reasoning through a chain of conditionals, each conditional of the form "if such-and-such is possible for x then so is such-and-such," in which the antecedent can have a degree of truth that is slightly higher than that of the consequent, because the antecedent mentions something that is more possible for x, relative to the way things actually are, than the consequent does. Such a gap prevents the premises which manifest it from being wholly true, so an argument that depends on (4) in this way is unsound.

An alternative way of implementing the degree-of-truth idea is to use the counterpart-theoretic approach to the modeling of possibilities due to Lewis (1968; 1986, ch. 1). To model possibilities for a collection of objects $a_1, \ldots, a_n$ one does not configure those objects with properties and relations, but rather, for each self-contained possibility for $a_1, \ldots, a_n$ which one wishes to model, one selects objects $b_1, \ldots, b_q$ to be the counterparts or representatives of $a_1, \ldots, a_n$ in the model, and then configures those counterparts with properties and relations. Its being possible for a to be f is then modeled by a setup in which a has a counterpart b configured so that f is true of it (that is, of b).

This approach introduces certain degrees of freedom missing in the standard approach. For example, one of the a_i may have two or more representatives $b_{i1}, \ldots, b_{ij}$. Or two of the a_i may have the *same* counterpart b_j (thus, on this approach, we can make sense of the idea that actually distinct things could have been identical). However, more to the point in the present context is the fact that it makes good sense to treat the counterpart relation as a relation of degree. Repeating our previous example in the present context, we would have three possible situations, w, u, and v, such that it is part of w that x and $p_1, \ldots, p_{20}$ stand in the "made of" relation, part of u that a counterpart y of x and counterparts of $p_1, \ldots, p_{11}, p_{32}, \ldots, p_{40}$ stand in the "made of" relation, and part of v that a counterpart z of x and counterparts of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$ stand in the "made of" relation. But while y and z can be counterparts of each other to the maximum degree, x can only be represented by either to a much lesser degree, because of the great difference in make-up. Still, y can qualify as a slightly stronger counterpart of x than z is, because y has more than half its parts in common with x while z only has half in common.

Returning to (4), we now read it as asserting that if there is a world where a counterpart of x has such-and-such a make-up, then there is a world where a counterpart of x has a slightly different make-up. Definition (4) generates such conditionals as

(8) If it is possible for x to be made of $p_1, \ldots, p_{11}, p_{32}, \ldots, p_{40}$, then it is possible for x to be made of $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$

which we interpret counterpart-theoretically as

(9) If there is a world where x has a counterpart made of (counterparts of) $p_1, \ldots, p_{11}, p_{32}, \ldots, p_{40}$, then there is a world where x has a counterpart made of (counterparts of) $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$.

But (9) has an antecedent "there is a world where x has a counterpart made of (counterparts of) $p_1, \ldots, p_{11}, p_{32}, \ldots, p_{40}$," which has a higher degree of truth than its consequent, "there is a world where x has a counterpart made of (counterparts of) $p_1, \ldots, p_{10}, p_{31}, \ldots, p_{40}$," because of the higher degree of counterparthood associated with the constitution closer to that of x. Again, therefore, the argument against the sufficiency of intrinsic and spatio-temporal duplication for identity that (4) seemed to support is shown to be unsound. Being a universal statement, (4) itself cannot be any more true than the least true of its instances. So on both the counterpart-theoretic and the relative possibility accounts, (4) comes out as at best slightly less than wholly true.

In passing, let us note that this apparatus solves the problem with which we ended §2, the problem of how to make sense of the idea that an individual's natural essence might be vague. In the jargon of the counterpart-theoretic approach, we can understand the claim that it is part of the natural essence of x to be made of a *substantial proportion of* $p_1, \ldots, p_{20}$ as allowing that an object y which at another world w is made of counterparts of a substantial proportion of $p_1, \ldots, p_{20}$ can be fully a counterpart of x, while an object z not meeting this condition cannot (the vagueness of individual essence can also be accommodated by the relative possibility approach). Naturally, the solution generalizes to other respects in which things could have differed from the way they actually are to various degrees. For example, the previous considerations can be repeated with respect to the design of an artifact.

It is not appropriate here to attempt to adjudicate between the relative possibility and counterpart-theoretic diagnoses of the flaw in modal slippery-slope arguments, since this would take us rather far afield into some of the technical esoterica of modal logic and metaphysics (see Ramachandran, 1989). The point is simply that on both accounts, the argument against duplication as sufficient for identity is plausibly convicted of the same fallaciousness as affects the better-known sorites paradoxes which do not involve modality. Thus, while we have not gone beyond intuition ourselves yet in defense of the sufficiency of duplication for identity across worlds, at least we have shown that the intuition is consistent with others that also have some force.

A different kind of objection to intrinsic and spatio-temporal duplication as sufficient for identity has been given by T. J. McKay (1986). McKay's example is a direct counterexample to the sufficiency condition that does not rely on any explicit supplementary argumentation for its force. Suppose we have an organism o_1 arising from a cell c. Then, even though highly improbable, it is conceivable that one way or another the molecules composing c should eventually detach themselves from o_1 and reassemble to constitute a cell c' in which the same molecules are configured in the same way as in c. We shall suppose that $c' = c$, since if this is denied, c' and c may themselves be used in a McKay-style counterexample to the sufficiency of intrinsic and spatio-temporal duplication for identity across worlds. So in this story, what happens is that c reassembles.

But if c reassembles having given rise to o_1, we may suppose that c now gives rise to an organism o_2 by the same biological processes through which it gave rise to o_1. And so long as it does this while o_1 still exists, there is no doubt that $o_1 \neq o_2$. Indeed, logically we can have this phenomenon repeating itself as rapidly as we like to give as many organisms as we like, but two suffice to make the point. Organisms o_1 and o_2 are otherwise unexceptional, and so they both have the usual mundane range of possibilities open to them, concerning when they come into existence, what routes they trace through space, what food they ingest, what fate ultimately befalls them, and so on. Indeed, for any life which o_1 might have had, it seems

that o_2 might have had the same life. Only possibilities which somehow involve one of the o_i rather than the other would be unshareable (for example, o_1 can exist in a world where o_2 does not exist, but o_2 cannot exist in such a world). However, if that is as much as the modal differences between the two amount to, then there are worlds *u* and *v* such that in *u*, o_1 leads such-and-such a life, and neither o_2 nor anything else originating from *c* exists; while in *v*, o_2 leads the same life, while neither o_1 nor anything else originating from *c* exists. Since there need be no other differences in context, it seems that the only required difference between *u* and *v* is in the mere identity of a certain organism. Thus, as it is in *u*, o_1 is intrinsically and spatio-temporally indistinguishable from o_2 as the latter is in *v*. But o_1 and o_2 are different things, hence the counter-example to the sufficiency of intrinsic and spatio-temporal duplication for identity across worlds.

The case has undeniable force. If we can correctly say "*this* one might have been thus-and-so" pointing at o_1, why cannot we equally well say "*that* one might also have been thus-and-so," pointing at o_2, when it is perfectly ordinary possibilities for organisms which fill out the "thus-and-so"? On the other hand, the idea that there are such worlds as *u* and *v* has the counter-intuitive consequence that the normal biological creation processes by which an o_i comes into existence do not settle *which* of the o_i it is which is coming into existence: some extra ingredient, the identity of the organism itself, remains to be added. In a well-known paper on personal identity, Chisholm propounded the view that if functionally equivalent hemispheres of a human brain were transplanted into separate bodies in such a way that no relation not presupposing identity distinguishes one of the new individuals from the other in terms of how he stands to the original owner of the brain ("Oldman"), there might nevertheless be a fact of the matter to the effect that one rather than the other of the new individuals is identical to Oldman (Chisholm, 1970). Since Chisholm is not positing anything like Cartesian substance, the idea that such an identity may hold in the absence of any distinguishing feature of the sort normally taken to be relevant to identity is difficult to understand. Similarly, confronted with o_1 and o_2, the idea that one rather than the other of these is identical to an organism which could have come into existence had things been thus-and-so is equally puzzling. To be able to point at o_1 and o_2 and intone "there could have been an organism which is thus-and-so and is identical to *this* organism, not that one" makes the consequences no more intelligible than they are in Chisholm's case if we point at one of the new individuals and say "*This* one, not *that* one, was Oldman."

However, a significant difference between the modal case and the split-brain case is that in the former there is an inevitable asymmetry to which appeal can be made to justify exclusion of *v*. For in the actual world, o_2 is the second organism to originate from *c*, while in *v* it is the first, since it is stipulated to be the only organism which develops in *v* from *c*. The untoward consequences of admitting *v* may then be taken as an argument for its being *essential* to o_2 to be the second organism to originate from *c* rather than the first, or a later one, and *essential* to o_1 to be the first. We will call this the *essentiality of order*. Then if we count the fact that o_2 is the second to originate from *c* as being intrinsic to o_2, we can still have intrinsic and spatio-temporal duplication across worlds as a sufficient condition for identity.

"Intrinsic" is a term of art, and we are free to extend it in this way, but only if we apply it consistently, counting other properties relevantly like *being the second organism to develop from c* as intrinsic too. Yet earlier, we defended the essentiality of origin in part by pointing out that some forms of skepticism about it allow the identity of an organism which develops from a certain cell at a world to turn on whether or not some *other* cell gives rise to an

organism at that world. So we do not want to say that being in a world where such-and-such another cell does not give rise to an organism is intrinsic to a given organism. How, then, is it any better to allow the identity of an organism which develops from *c* at a world to turn on whether or not *c* has *already* given rise to an organism at that world?

There are two respects in which the cases differ. First, the proposal to allow the identity of the organism which develops from c_1 at *w* to turn on whether any organism develops from c_2 at *w* has consequences that the essentiality of order does not. The main one is that in *w*, on the former proposal, future facts play a role in fixing the identity of presently existing organisms, that is, facts about whether or not c_2 will give rise to an organism. And this seems wrong. Still, this only means that allowing the identity of an organism which develops from *c* to turn on whether it is the first, second, third, and so on to do so is not vulnerable to one particular objection; it does not make the classification of the order-property as intrinsic any more reasonable. However, on the positive side, we can note that there is a particularly intimate relationship between an organism and the cell from which it develops. The two cannot be said to be identical, since the cell ceases to exist when it divides while the organism is just beginning its existence. But there is another relation, variously known as temporary identity, or *coincidence*, or realization, which characterizes how the cell and the organism to which it gives rise stand to each other (Yablo, 1987). If having been coincident with a certain cell is an intrinsic feature of an organism at all, it seems a reasonable extension of the notion of intrinsicness to say that having been identical with a new, or an n-times-used, cell, is also intrinsic.

A move like this is partly terminological, designed mainly to keep the formulation of the sufficient condition for transworld identity simple: intrinsic and spatio-temporal duplication. The substantial issue is whether counting the order property as essential is justifiable; for if it is not, then arguments for the essentiality of a set's members, or an artifact's composition, or an organism's origin, are in trouble, since these apparently depend on constraints on identity that McKay's example, if conceded, would show to be incorrect. Yet these essentialist theses also have intuitive force. So there is an unavoidable trade-off here, unless we find a different way of justifying the essentialist theses.

4 The Grounds of Metaphysical Necessity

Perhaps the most difficult question surrounding the topic of essentialism is the problem of how the kinds of necessities we have been discussing arise. It seems, on the face of it, that these necessities require us completely to rethink our received notions about the grounds of necessity. According to the traditional view, which receives its paradigm formulation in Hume's writings, there are no "necessary connections" between distinct existences. But this is inconsistent with a thesis such as the essentiality of origin, which postulates a connection between an organism and the cell from which it arose that is necessary *modulo* existence of the two entities. On the positive side, according to Hume, whatever necessities there are, are to be explained in terms of "relations of ideas." But again, this is inconsistent with such a thesis as that Wiggins is essentially human: even if a name such as "Wiggins" has a meaning, grasp of which is a condition for mastery of the name, it is not plausible that "human" would be part of the meaning of a name of a human. For example, in a society with advanced robotic or biotechnology, one might have complete mastery of an individual's name without knowing whether or not that individual is human. And it is completely implausible that

some way of identifying the cell from which a named organism develops should be part of the content of any name of that organism. The overall problem for the traditional view is that the sorts of examples we have been discussing involve necessary truths which are *a posteriori*, while truths that are based on relations of ideas are supposed to be *a priori*.

However, our discussion in fact contains an implicit defense of the traditional association between the necessary and the *a priori*. For any specific essentialist thesis about a particular entity or entities, such as that Wiggins is human, is derived from two premises, one of which gives some *a posteriori* information about the relevant object or objects (such as that Wiggins *is* human) and the other of which states a modal principle to which objects of that sort must conform, such as the principle that the biological kind of an organism is essential to it. And the latter principle is *a priori* true if it is true at all. So the source of necessity may be in "relations of ideas" after all. This division of labor in the production of the necessary *a posteriori* renders essential properties a little less mysterious. The explanation of why some property of an object is temporary, or permanent, may have to do with the physical nature of the environment in which the object is situated, or with other physical features of the object; and it is hard to see how the explanation of why a property is accidental rather than essential or vice versa could make headway from similar resources. But it will not have to, if the status of a property as accidental or essential is settled by *a priori* principles.

On the other hand, for every alleged *a priori* truth which our defenses of certain essentialist theses have appealed to, there is a good question about the source of the truth of that principle. It would be convenient if the duplication-implies-identity principle or the no-bare-particulars principle could be accounted for as the product of definitions or decisions ("conventions"). But it is unsatisfactory to suppose that these truths are manufactured by stipulations, whether explicit or implicit. To the extent that one finds the principles plausible, they seem forced on us by the nature of our concepts. At this point, it looks as if the best one could hope for is a Strawsonian "descriptive metaphysics" which spells out the aspects of our concepts that account for the force of the principles.

It is surely no accident that essentialist theses are asserted for unified classes of things. We do not find any grounds for holding that there are *some* sets whose members belong to them accidentally, while others have their members essentially, or *some* organisms whose origin is essential to them and others whose origin is accidental. A hypothesis which would explain this is that we have a general conception of what a set is, or what an organism is, and the conception of a specific set, or a specific organism, as an instantiation of the general conception. That is, the general conception of a set, or an organism, has certain *parameters*, and our understanding of what is involved in the existence of a particular set, or organism, is simply that these parameters take on particular values: no less, *and* no more, is involved. In more detail, an organism is conceived of as a thing with a particular origin in some reproductive process, a particular nature deriving from the entities involved in that process, and a subsequent career that traces some continuous spatio-temporal path through the world; and any particular organism is simply a particular manner of instantiation of these parameters. There is, of course, no simple step from parameter-instantiation to essential property, as the inclusion of spatio-temporal path here illustrates. The point is, rather, that if there is no more to individuality than instantiation of the parameters of a general conception, then there is no sense to a notion of identity which transcends instantiation of parameters. But it seems that opposition to essentialist theses of the principles on which, we have argued, they rest inevitably leads to such a notion of identity. For example, on McKay's view,

the parameters of the general conception of organism may be instantiated at a world without this being sufficient to fix which organism it is that results from the instantiations. And on a view that denies that the kind of a thing is essential to it, we arrive at a notion of bare particular, a thing which has a specific identity without, it seems, instantiating the parameters of any general conception at all.

The proposed scheme, then, if it is an accurate reflection of our concepts, offers some justification for some essentialist theses. What other kinds of justification are there? In §1 of this chapter, we noted some technical difficulties in getting the definitions of "essential property" and "individual essence" exactly right. Fine has suggested that these problems are symptomatic of a deeper inadequacy in an approach to these concepts which tries to explain them in terms of what is possible and what is necessary (Fine, 1994). According to Fine, a truth of the form "P is essential to x" will certainly give rise to modal truths, such as that necessarily, if x exists then x has P. But simply because facts about what is essential give rise to such truths, it does not follow that essentialist concepts are modally explicable. And in fact they are not, he argues, since it is possible for the same range of modal facts to be determined by different, incompatible, collections of essentialist facts. For example, at least three competing views of the essentialist facts about persons, bodies, and minds might give rise to the same range of modal facts, the three views disagreeing over how a person, her body, and her mind are related. If a person p has a body b and a mind m, it may be true that, necessarily, if p exists p has b and m; but one philosopher might say that this is because persons are fundamental, though essentially possessing the bodies and minds they do, while another might say that this is because human bodies are fundamental, giving rise to persons when they (the bodies) realize sentience, but giving rise to the same mind and therefore the same person in any possible situation where they give rise to a mind. What this brings out, Fine says, is that an essentialist truth has its source in specific objects, but the modal truths to which essentialist truths give rise do not determine which particular objects are the sources of the essentialist truths.

What Fine is suggesting, then, is that modal concepts are insensitive to essentialist facts in much the same way as they are to intensional facts. It is familiar that analogs of intensional notions like proposition and belief can be defined in modal terms, using the possible-worlds apparatus; but if these analogs are taken to be the intensional notions themselves, there are various counter-intuitive consequences, such as that everyone believes all the logical consequences of their beliefs, and that there is only one necessarily true proposition. In a similar way, our definition (3), for example, has the consequence that an object can have more than one individual essence.

In place of the modal approach, Fine would put the Lockean (or Aristotelian) idea of *real definition* in center stage for the explanation of essentialist concepts. We commonly think that it is words, or concepts, which admit of definition; but if we can make sense of ordinary things being definable as well, then the individual essence of a thing would be exactly what that thing's definition delivers. Correspondingly, essential properties would be those which flow from individual essence, and there would be no reason to expect existence, or being a member of $\{x\}$, to be essential to an ordinary material object x. With these ideas worked out in detail, we have a rival to the modal approach, and the project of resolving the problems with the modal definitions, if they can be resolved, takes on more urgency. So despite the fact that investigation of essence stretches back at least to Aristotle, we can expect lively and ongoing research to extend it; it is yet another subject whose history is definitely not at an end.

References

Aristotle. 1928. "Topica." In *The Works of Aristotle*, vol. 1, edited by Sir David Ross, translated by W. A. Pickard-Cambridge. Oxford: Oxford University Press.

Chisholm, R. 1967. "Identity through possible worlds: some questions." *Noûs*, 1(1): 1–8.

Chisholm, R. 1970. "Identity through time." In *Language, Belief and Metaphysics*, edited by H. E. Kiefer and M. Munitz, pp. 163–182. New York: State University of New York Press.

Chomsky, N. 1988. *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

Fine, K. 1977. "Postscript." In *Worlds, Times and Selves*, A. N. Prior and K. Fine, pp. 116–168. London: Duckworth.

Fine, K. 1981. "First-order modal theories I – sets." *Noûs*, 15(2): 177–205.

Fine, K. 1994. "Essence and modality." *Philosophical Perspectives*, 8: 1–8.

Kripke, S. 1972. "Naming and necessity." In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 252–355. Dordrecht, Netherlands: Reidel.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.

Lewis, D. 1968. "Counterpart theory and quantified modal logic." *Journal of Philosophy*, 65(5): 113–126.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

McGinn, C. 1976. "On the necessity of origin." *Journal of Philosophy*, 73(5): 127–135.

McKay, T. J. 1986. "Against constitutional sufficiency principles." In *Midwest Studies in Philosophy XI: Studies in Essentialism*, edited by P. A. French, T. E. Uehling, and H. K. Wettstein, pp. 295–304. Minneapolis: University of Minnesota Press.

Putnam, H. 1975. "The meaning of 'meaning.'" In *Philosophical Papers, vol. 2, Mind, Language and Reality*, pp. 215–271. Cambridge: Cambridge University Press.

Ramachandran, M. 1989. "An alternative translation scheme for counterpart theory." *Analysis*, 49(3): 131–141.

Salmon, N. 1986. "Modal paradox: parts and counterparts, points and counterpoints." In *Midwest Studies in Philosophy XI: Studies in Essentialism*, edited by P. A. French, T. E. Uehling, and H. K. Wettstein, pp. 75–120. Minneapolis: University of Minnesota Press.

Salmon, N. 1989. "The logic of what might have been." *Philosophical Review*, 98(1): 3–34.

Wiggins, D. 1980. *Sameness and Substance*. Oxford: Blackwell.

Yablo, S. 1987. "Identity, essence, and indiscernibility." *Journal of Philosophy*, 84(6): 293–314.

Further Reading

Forbes, G. 1985. *The Metaphysics of Modality*. Oxford: Clarendon Press.

Mackie, J. L. 1974. "*De* what *re* is *de re* modality?" *Journal of Philosophy*, 71(16): 551–561.

Plantinga, A. 1974. *The Nature of Necessity*. Oxford: Clarendon Press.

Postscript

PENELOPE MACKIE

In the main text, it is pointed out that two theses about the essential properties of individuals seem intuitively very plausible: the thesis – known as the essentiality (or necessity) of origin – that individuals such as organisms and artifacts have certain particular features of their origins essentially, and the thesis that the broad kind to which something belongs is essential to it. As is also pointed out in the text, it is one thing to find such theses intuitively

plausible, and another to discover principles that explain why certain properties are essential properties. This postscript explores some of these issues in more detail.

The Essentiality of Origin and Individual Essences

The argument given in the text for the essentiality of origin is primarily an argument for what I shall call *substantial individual essences* – individual essences that consist in intrinsic features of a thing other than the property of identity with that thing. A substantial individual essence would thus provide a non-trivial answer to the question what *makes* something the particular object that it is – what being that particular object consists in. In the main text it is argued that, if things such as organisms and artifacts have substantial individual essences, they include certain unique features of the way those things originated, thus supporting a version of the essentiality of origin (§2).

As indicated in §1, a thing's individual essence may be defined as a property, or set of properties, that is necessarily both necessary and sufficient for being that thing. (According to the characterization (3) in §1, if e is an individual essence of x, then x has e essentially, and anything, in any possible world, that has e is identical with x. But if anything in any possible world that has e is identical with x, having e is necessarily sufficient for being x.)

The "sufficiency" aspect of an individual essence has important implications. If an object has a substantial individual essence, its essential properties must include at least one that is immune to duplication: a property that is "unshareable" in the sense that it cannot be possessed by two objects in any possible world. Unless the set includes such a property, possession of the set of properties cannot, strictly, be sufficient for identity with the individual in question. In the text, it is suggested that if certain unique features of the way that a particular thing originated are essential to it, this demand for "unshareable essential properties" can be met. Hence this individual-essence-based argument may be regarded as a version of a "sufficiency" argument for origin essentialism (Robertson and Atkins, 2013).

In the main text, it is argued that substantial individual essences are required to rule out the possibility of duplication without identity – more precisely, to rule out there being pairs of possible worlds containing objects that are intrinsic and spatio-temporal duplicates and yet numerically distinct. Thus the argument for substantial individual essences, and hence the associated argument for origin essentialism, depends crucially on the assumption that such duplication without identity is unacceptable. As is noted, however, this assumption may be questioned.

One ground for questioning it (§3) is a variant of Chisholm's paradox: an argument that appears to lead, from the apparently uncontroversial assumption that a thing's origin could have been slightly different from the way that it actually was, to the conclusion (via the transitivity of identity) that there may be two possible worlds containing entities that are distinct although they are intrinsic and spatio-temporal duplicates. As is pointed out in the text, however, suspicion is cast on the soundness of this argument by its similarity to a standard sorites paradox of vagueness. It is therefore plausible to suppose that there is some flaw in the Chisholm's paradox argument, and several diagnoses are explored in the main text. Moreover, as noted there, diagnosing a flaw in the Chisholm's paradox reasoning is of significance for the justification of origin essentialism independently of the defense of the "no duplication without identity" principle. For the only intuitively plausible versions of the essentiality of origin combine the view that an artifact (such as a watch, say) could *not* have

come into existence from completely different matter with the concession that the artifact *could* have come into existence from slightly different matter. The Chisholm's paradox reasoning, if sound, would show that this combination of "modal restriction" with "modal tolerance" is incoherent.

As is acknowledged in the main text, however, blocking the Chisholm's paradox argument does not solve all the problems for the individual-essence-based approach to origin essentialism. A second reason for questioning the crucial prohibition on duplication without identity has its roots in a problem, identified by McKay (1986), that has become known as the "recycling problem." McKay's argument shows that a property such as *being an organism of kind K that develops from cell c*, even if it is in fact possessed by only one organism, and even if it is an essential property of that organism, cannot be a sufficient condition for identity with that organism. For as long as *c* could be "recycled" (to produce a second organism of kind *K* at a later time), the property could be duplicated – in the sense of being instantiated by more than one thing in a single possible world. Such a "recycling world" is not itself a counter-example to the "no duplication without identity" principle, since the two organisms in the recycling world have, in that world, different spatio-temporal properties. But unless there is some difference in the *essential* properties of the two organisms, there can be possible worlds that differ only in that one of them contains one of the organisms, whereas the other contains the other, in conflict with the "no duplication without identity" principle (§3).

The response proposed in the text to the recycling problem is to protect the view that persisting things such as organisms have substantial individual essences – and the associated "no duplication without identity" principle – by adding a principle of the *essentiality of order*. Evidently, if it is essential to an organism o_1 – a particular tree, say – that it is not only a tree that develops from cell *c*, but also the *first* tree that develops from *c*, then the tree has, after all, an essential property that is immune to duplication, and distinguishes it, in any possible world, from any other tree that develops from cell *c*.

However, this solution comes at a price. For one thing (as noted in the main text) it is not obvious that a property such as being the first organism of its kind to develop from cell *c* is an intrinsic property. Second, a conception of individual essences that incorporates the essentiality of order appears to have counter-intuitive consequences. For example, if it is not only necessary, but also (necessarily) sufficient, for identity with an organism that was in fact the only tree to originate from cell *c*, to be the *first* tree to originate from cell *c*, then in a "recycling" world in which *c* generates a tree in the year 1500, and a second tree in the year 2010, the first tree, and not the second, must be identified with the original tree, even if the 2010 tree in the recycling world is otherwise an exact duplicate of the actual tree in every respect, including the time of its origin (cf. Hawthorne and Gendler, 2000; Forbes, 2002).

Such difficulties do not refute the individual-essence-based argument for the essentiality of origin. Nevertheless, they are serious enough to cast some doubt on the thesis that persisting things really do have substantial individual essences, and on any argument for origin essentialism that requires this assumption.

Faced with this problem, one might seek to defend origin essentialism without appealing to individual essences. One initially attractive line of thought is prompted by the observation that, when we consider ways in which persisting individuals could have been different, we typically focus on ways in which they could have *become* different. If *every* possibility for an actual individual such as an organism or an artifact has to be a way that it could have

developed from the way that it actually was at some time in its existence, then no such individual could have had a different origin. For once it has come into existence, it is, of course, too late for it to acquire a different origin. However, this line of argument is highly problematic as a defense of the essentiality of origin thesis. It seems that it would prove far too much, since it would imply that an individual could not have had an origin different in any respect (even, for example, in its location) from its actual origin.

These and other difficulties have led some to conclude that there is no coherent justification for the essentiality of origin thesis, even if it may be possible to explain its intuitive appeal (Mackie, 2006, ch. 6). A further suggestion, made by David Lewis, is that the kernel of truth in the thesis is simply that in some contexts of counterfactual speculation about a thing it is appropriate to keep fixed certain features of its origin (1986, p. 252). Yet others have attempted to provide novel arguments for the thesis (Rohrbaugh and deRosset, 2004, and, for discussion, Cameron and Roca, 2006, and Forbes and Robertson, 2006). In spite of extensive discussion over several decades, there is no consensus on the status of the undeniably appealing, yet extremely perplexing, thesis of the essentiality of origin.

Finally, if things such as persons, organisms, and artifacts do not have substantial individual essences, what does their individuality consist in: what, for example, is it that makes Socrates the individual that he is and no other? One answer that has been proposed is that it is the possession of a *haecceity* – a *sui generis* non-qualitative essential property that is unique to the individual. Another answer is that there is *nothing* that makes Socrates the individual that he is, if what is at issue is whether there is any essential property that is non-trivially sufficient for identity with Socrates (Mackie, 2006). Rather confusingly, both views are referred to in the literature as varieties of "haecceitism" (see Mackie and Jago, 2013, §4).

The Essentiality of Kind Membership

Even if the view that ordinary individuals have substantial individual essences or essential origin properties cannot, in the end, be sustained, that would, of course, leave unscathed the second type of essentialism mentioned in the main text, according to which things belong essentially to certain sorts or kinds. However, it is one thing to say that the broad kind to which a thing belongs is an essential feature of it, and another to decide which are the "essential kinds."

One relatively modest thesis is that everything belongs essentially to some general ontological category, such as *substance*, *event*, *property*, and so on. This "category essentialism" would apparently explain why David Wiggins could not have been a musical performance, for example. If, however, what is sought is an explanation (or justification) of claims such as the claim that Wiggins could not have been a lake or a polar bear or a forest, or that a pot of marmalade could not have been a railway station (§2), evidently some more restrictive notion of essential kinds is required.

A popular idea is to invoke the notion, which has its roots in Aristotle's theory of essence and accident, of a *substance sortal* (or *substance concept*). A substance sortal is a concept that plays a special role in individuation: it not only provides a criterion of identity, but also represents a "necessarily permanent" property: one that applies to an object throughout its existence if it applies to it at any time in its existence, and hence is a property that the object cannot lose without ceasing to exist (Wiggins, 1980, ch. 3; 2001, ch. 3). According to this criterion, it is plausible to regard *cat* as a substance sortal (nothing can change from being a

cat to not being a cat, or vice versa, without going out of existence), although *kitten*, while evidently in some sense an individuative concept, is not a substance sortal.

However, it is one thing to say that an object cannot change from being a cat to not being a cat, or vice versa, without going out of existence, and another to say that an object that is a cat could not have existed without ever being a cat. Being a necessarily permanent property does not entail being an essential property, as is demonstrated by properties such as *originating in Paris*, or *being at all times unmarried*.

In response, some defenders of a "sortal-based" account of essential kinds will argue, as E. J. Lowe has done, that, in the case of substance sortal properties, the reason *why* the property is necessarily permanent is that it is also an essential property. For example, if the cat Tibbles cannot cease to be cat without ceasing to exist, perhaps the reason is that "a cat" specifies *what Tibbles is*, in a sense that implies that *being a cat* is at least part of her essence (Lowe, 2007). According to this argument, all substance sortals would represent essential properties.

A different way of developing a sortal-based account of essential kinds is suggested by David Wiggins, who argues that it is only *ultimate sortals* that represent essential properties, where an ultimate sortal is the most general sortal corresponding to a given principle of individuation (Wiggins, 1980; 2001). Since, on Wiggins's theory, two substance sortals may provide the same principle of individuation, this version of a sortal-based account of essential kinds yields different – and conflicting – results from those implied by Lowe's proposal.

Essential Properties and What a Thing Is

It is worth noting that the proposal just considered – that the Aristotelian notion of a substance sortal is the key to understanding which kind properties are essential – has clear affinities with Fine's (1994) suggestion that the notion of essential property that is of interest to metaphysics is tied to the notion of a thing's nature as revealed by a real definition that says "what the thing is." It is a further, interesting, question whether (as is suggested in the main text) a conception of essence in terms of real definition requires that everything has a unique real definition (and thus has either its own substantial individual essence or its own haecceity), or whether, for example, the real definition of Socrates might be the very same as the real definition of Plato.

It can also be noted that although Fine's proposed counter-examples to the standard (modal) conception of essence in terms of necessary properties (and his associated claim that the modal conception goes wrong in including, among a thing's essential properties, necessary properties of the thing that do not flow from its nature or essence) have been widely accepted, not all have agreed with Fine in drawing the moral that metaphysical modality should be understood in terms of a prior notion of essence. In particular, there have been attempts to provide a revised modal theory of essence that characterizes a thing's essential properties as a proper subset of its necessary properties demarcated by a restriction that does not treat the notion of nature or essence as primitive or fundamental (Cowling, 2013; Wildman, 2013; and, for discussion, Skiles, 2015).

Altogether, then, questions about the ultimate source of essential properties, and about what essential properties things have, remain live issues that are the subject of ongoing and unresolved debate.

References

Cameron, R., and S. Roca. 2006. "Rohrbaugh and deRosset on the necessity of origin." *Mind*, 115(458): 361–366.

Cowling, S. 2013. "The modal view of essence." *Canadian Journal of Philosophy*, 43(2): 248–266.

Fine, K. 1994. "Essence and modality." *Philosophical Perspectives*, 8: 1–8.

Forbes, G. 2002. "Origins and identities." In *Individuals, Essence and Identity: Themes of Analytic Metaphysics*, edited by A. Bottani, M. Carrara, and P. Giaretta, pp. 319–340. Dordrecht, Netherlands: Kluwer.

Forbes, G., and T. Robertson. 2006. "Does the new route reach its destination?" *Mind*, 115(458): 367–374.

Hawthorne, J., and T. S. Gendler. 2000. "Origin essentialism: the arguments reconsidered." *Mind*, 109(434): 285–298.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lowe, E. J. 2007. "Review of Mackie, *How Things Might Have Been*." *Mind*, 116(463): 762–766.

Mackie, P. 2006. *How Things Might Have Been: Individuals, Kinds, and Essential Properties*. Oxford: Clarendon Press.

Mackie, P., and M. Jago. 2013. "Transworld identity." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/fall2013/entries/identity-transworld/ (accessed August 27, 2016).

McKay, T. J. 1986. "Against constitutional sufficiency principles." In *Midwest Studies in Philosophy XI: Studies in Essentialism*, edited by P. A. French, T. E. Uehling, and H. K. Wettstein, pp. 295–304. Minneapolis: University of Minnesota Press.

Robertson, T., and P. Atkins. 2013. "Essential vs. accidental properties." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/win2013/entries/essential-accidental/ (accessed August 27, 2016).

Rohrbaugh, G., and L. deRosset. 2004. "A new route to the necessity of origin." *Mind*, 113(452): 705–725.

Skiles, A. 2015. "Essence in abundance." *Canadian Journal of Philosophy*, 45(1): 100–112.

Wiggins, D. 1980. *Sameness and Substance*. Oxford: Blackwell.

Wiggins, D. 2001. *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Wildman, N. 2013. "Modality, sparsity, and essence." *Philosophical Quarterly*, 63(253): 760–782.

Further Reading

Adams, R. M. 1979. "Primitive thisness and primitive identity." *Journal of Philosophy*, 76(1): 5–26.

Hale, B. 2013. *Necessary Beings: An Essay on Ontology, Modality, and the Relations Between Them*. Oxford: Oxford University Press.

Salmon, N. 2005. *Reference and Essence*, 2nd edn, with added appendices. Amherst, NY: Prometheus Books.

35

Reference and Necessity

ROBERT STALNAKER

Saul Kripke remarked, at the beginning of his lectures, *Naming and Necessity* (1980), that he hoped his audience would see some connection between the two topics mentioned in his title. In those lectures Kripke defended some bold theses, some about naming that belong to semantics and the philosophy of language, others about necessity that belong to metaphysics. It is clear that the arguments for the different theses were interrelated, but it remains a matter of debate just what the connections are, both in Kripke's argumentative strategy, and in the issues themselves. Kripke and Hilary Putnam were criticized for attempting to derive metaphysical conclusions – about the essential properties of things – from premises in the philosophy of language about the nature of reference and the semantics of proper names. One might instead think that the direction of Kripke's arguments go the other way: that conclusions about reference and proper names were derived in part from controversial metaphysical assumptions about possible worlds and essential properties. Either way, there is reason to be puzzled: on the one hand, one might be skeptical (to borrow the metaphor that Nathan Salmon used to express his puzzlement about this) that one could, without sleight of hand, pull a metaphysical rabbit out of a linguistic hat (see Salmon, 1981). On the other hand, one might wonder why a proper understanding of the way our language happens to work should require controversial assumptions about the metaphysical nature of the world that our language talks about. My aim in this chapter is to try to resolve some of this puzzlement by clarifying the relationship between theses and questions about reference and theses and questions about necessity and possibility. In the background of my discussion will be very general questions concerning how claims about the way we talk about the world relate to claims about what the world must be like, but in the foreground will be more specific questions concerning the relations between the different theses Kripke defends about individuals and their names. My main claim will be that Kripke's contribution was not to connect metaphysical and semantic issues, but to separate them: to provide a context in which questions about essences of things could be posed independently of assumptions about the semantic rules for the expressions used to refer to the things, and in which

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

questions about how names refer could be addressed without making assumptions about the nature of the things referred to. I will argue that Kripke's theses about proper names and reference do not presuppose any metaphysical theses that ought to be controversial, though even stating those theses requires a framework that might be thought not to be metaphysically neutral. And I will argue that no metaphysical conclusions are derived from theses about reference and names, although clarification of the nature of reference helps in the rebuttals to arguments against metaphysical theses that Kripke defends.

I will start in §1 by contrasting three kinds of questions that Kripke discusses in *Naming and Necessity* – two that belong to semantics and the philosophy of language, and one that belongs to metaphysics – and sketching the answers that Kripke defends, along with contrasting answers that he criticizes. Then, in §2, I will discuss the apparatus that he uses to clarify his questions – the possible-worlds framework – and argue that it should be understood not as a metaphysical theory, but as a methodological framework in which alternative metaphysical and semantic theses can be stated. In §§3–5 I will look in more detail at the arguments for the different theses and the way the three different kinds of issues interact. (See also Chapter 34, ESSENTIALISM; Chapter 36, NAMES AND RIGID DESIGNATION; and Chapter 31, MODALITY.)

1 Questions and Theses

At this point my aim is just to set the stage by making some simple distinctions between questions, and stating, without much explanation or argument, some alternative answers to the questions. First there are questions of what I will call "descriptive semantics." A descriptive-semantic theory is a theory that says what the semantics for the language is, without saying what it is about the practice of using that language that explains why that semantics is the right one. A descriptive-semantic theory assigns *semantic values* to the expressions of the language, and explains how the semantic values of the complex expressions are a function of the semantic values of their parts. The term 'semantic value,' as I am using it, is a general and neutral term for whatever it is that a semantic theory associates with the expressions of the language it interprets: the things that, according to the semantics, provide the interpretations of simple expressions, and are the arguments and values of the functions defined by the compositional rules that interpret the complex expressions. If, for example, the semantic theory in question assigns senses or intensions to the names and predicates of a language, and explains the senses of complex expressions as a function of the senses or intensions of their parts, then as I am using the term, the semantic values of that semantic theory will be the senses or intensions. The particular descriptive semantic question we will be concerned with is the question, what kind of thing is the semantic value of a proper name?

Second, there are questions, which I will call questions of 'foundational semantics,' about what the facts are that give expressions their semantic values, or more generally, about what makes it the case that the language spoken by a particular individual or community has a particular descriptive semantics. The specific question of this kind that we will focus on is the question, what is it about the situation, behavior, or mental states of a speaker that makes it the case that a particular proper name, as used by that speaker in a particular linguistic community, has the semantic value that it has? (See also Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §2; Chapter 8, A GUIDE TO NATURALIZING SEMANTICS.)

Third, there are questions about the capacities and potentialities of the things in the domain forming the subject-matter of some language; what, for example, might have been true of the things, such as persons and physical objects, that are the referents of some particular proper names?

Kripke's answer to the first question – the descriptive-semantic question about proper names – is the Millian answer: the semantic value of a name is simply its referent. The contrasting answer that he argued against is that the semantic value of a name is a general concept that mediates between a name and its referent: a concept of the kind that might be expressed by a definite description. According to this contrasting answer, the semantic value of the name – its sense or connotation – determines a referent for the name as a function of the facts: the referent, if there is one, is the unique individual that fits the concept, or perhaps the individual that best fits the concept.

Kripke's answer to the second question – the foundational-semantic question – is that a name has the referent that it has in virtue of a causal connection of a particular kind between the use of the name and the referent; the referent is the individual that plays the right role in the causal explanation of the fact that the name is being used, in the particular context in question, in the way that it is being used. In the case of this question, it is less clear what the contrasting thesis is, since the question is not explicitly addressed by the philosophers whom Kripke is criticizing. But what seems to be suggested is that the sense of a name is determined by the abilities and dispositions of the speaker to describe or identify a certain individual.

In response to the question about the capacities and potentialities of the things that we commonly refer to with names, Kripke defends the thesis that it makes sense to talk about the logical potential of an individual thing independently of how it is referred to, and that this potential is greater in certain ways, and less great in others, than some philosophers have supposed. For example, Shakespeare need not have been a playwright; he need not have written anything at all. He might have died in infancy, and been someone of whom we had never heard – even someone of whom all trace was lost after the seventeenth century. He might not have been called "Shakespeare," by us, or by anyone. His plays, or at least plays which are word-for-word just like his plays, really might have been written by someone else of the same name. On the other hand, Shakespeare could not possibly have been anything other than a human being, and he could not possibly have had parents other than the ones that he in fact had. In contrast, others have conceded that Shakespeare might have lacked any one of the attributes commonly attributed to him, but argued that he could not have lacked them all. Attributes not commonly known to apply to Shakespeare, such as having the particular parents that he in fact had, are all attributes that he might have lacked.

One thing I will argue is that while Kripke defends these theses about the descriptive semantics of names, the way the reference relation is determined, and the capacities and dispositions of human beings and physical objects (and I think he makes a persuasive case for each of them), his most important philosophical accomplishment is in the way he posed and clarified the questions, and not in the particular answers that he gave to them. I will suggest that we might buy Kripke's philosophical insights while rejecting all of the theses – while opting for a pure, Russellian description-theory of ordinary names, a non-causal account of the way names get associated with their values, and, in metaphysics, either for an anti-essentialist thesis according to which Shakespeare might have been a lamppost or a fried egg, or for a Leibnizian theory according to which Shakespeare had even his most apparently accidental properties essentially. The positive case for the theses that Kripke

defends is not novel philosophical insight and argument, but naïve common sense. The philosophical work is done by diagnosing equivocations in the philosophical arguments for theses that conflict with naïve common sense, by making the distinctions that remove the obstacles to believing what it seems intuitively most natural to believe.

2 The Possible-Worlds Framework

To accept what I will argue is Kripke's main philosophical contribution, you do have to buy a framework, an apparatus that he used to sharpen and clarify the contrasting theses, both semantic and metaphysical. I won't try to claim that the apparatus is either semantically or metaphysically neutral, but I will argue that the motivation and commitments of the framework are more methodological and conceptual than they are metaphysical. Philosophers often talk as if the decision to theorize with the help of this framework is, if one takes the claims one makes while using it seriously, a specific ontological commitment to a certain kind of entity. Some philosophers reject the framework because they reject the ontological commitment that its serious use makes. One hears, "I don't believe in possible worlds," as one might hear people say that they don't believe in transubstantiation, or flying saucers from other planets (commitments that some philosophers believe are about as plausible as the commitment to possible worlds). I think this attitude is based on a misconception (although I have to concede that it is a misconception that some defenders of possible worlds share with the critics). It is not that it is a misconception to think that serious talk about possibilities commits one to the existence of the possibilities one claims there are, just as it is not a misconception to think that the literal use of quantifiers commits one to the existence of things that one purports to quantify over. But it is not the framework itself that makes the specific commitments, just as it is not the semantics for first-order logic that makes any particular ontological commitment. Suppose someone were to reject the standard (extensional) semantics for first-order logic on the ground that he did not believe in individuals, to which that semantics is ontologically committed. The proper response would be to point out that first-order semantics is a framework for doing ontology, and not a particular thesis about what ontology is correct. Individuals are whatever there is to talk about; the semantic theory itself says nothing about what there is to talk about, and so makes no particular ontological commitments. Quine's slogan "To be is to be the value of a bound variable" is not an ontological thesis, but an attempt to promote a framework in which the ontological commitments of alternative philosophical and scientific theories can be stated without equivocation and compared.

The Leibnizian slogan "Necessity is truth in all possible worlds" should be understood in a similar spirit. This slogan and the possible-worlds framework that it presupposes should, I think, be understood not as an attempt to provide an ontological foundation for a reduction of modal notions, but as an attempt to formulate a theoretical language in which modal discourse can be regimented, its structure revealed, equivocation diagnosed and avoided. Modal discourse – speech that involves words such as "may," "might," and "must" – is notoriously complex and problematic, providing fertile ground for ambiguities, both ambiguities of scope, that arise because the semantic structure of modal statements is complicated and not simply reflected in surface syntax, and ambiguities that arise from alternative senses and context dependence of the modal words. Modal words interact with each other and with quantifiers, descriptions, temporal modifiers, and grammatical tense, aspect,

and mood in complicated ways that are difficult to sort out. Philosophical puzzles about, for example, necessary connection and counterfactual dependence, reference to non-existent things, capacities, and dispositions, the ability to do otherwise, the necessity of the past and the openness of the future, will presumably not all be dissolved simply by getting clear about modal discourse; but everyone should agree that clarifying the discourse in which such problems are posed is an essential first step. It is important to separate disagreements based on contrasting interpretations of the way the language works from those about the claims that the language is being used to state. Whatever one's metaphysical beliefs about the reality that modal discourse purports to describe, one should agree that it would be nice to have a language that is free of some of the ambiguities that infect modal discourse, and in which the claims made with modal words and constructions might be paraphrased; a language that uses only parts of discourse that are relatively clear and uncontroversial (the indicative mood and quantifiers), but that still has the expressive power to make claims about what might, would, or must be true. Achieving such clarification does not require a reductive analysis of modal to non-modal concepts, and so it is not required that a canonical language in which we do modal semantics be built on some pure, non-modal foundation, any more than formal languages designed to clarify quantification needed to be built on some pure, non-quantificational foundation (whatever that would be). What is needed is only the kind of opportunistic departure from ordinary language involved in the boot-strap operation that Quine called "regimentation." In the kind of regimentation Quine recommends, we begin with *ad hoc* paraphrases to remove ambiguity, we introduce variables to facilitate cross-reference, and we adopt a syntax in which quantifier scopes are reflected in a simple and systematic way in the order of the symbols and the placement of parentheses. "The artificial notation of logic," Quine remarked, "is itself explained, of course, in ordinary language" (1960, p. 159). Similarly, the primitive resources of the possible-worlds framework are explained in ordinary modal language, and the explanations will be intelligible only to one who understands at least some of that part of language. A modal skeptic who doubts that anything both meaningful and true is said in modal discourse will doubt both the value and the intelligibility of a framework in which that discourse is clarified. But it may be that the source of the skepticism is in the equivocations and unclarities that the framework helps to remove.

The general strategy is to find a part of our modal discourse that seems relatively free of the particular equivocations and unclarities that infect modal discourse generally, a part that might be developed and used to clarify the rest. We look for a way of making modal claims that uses paraphrases to avoid problematic constructions, a way that uses forms of expression that may perhaps be stilted and less idiomatic than the familiar ones, but that will still be recognizable paraphrases of ordinary modal claims. The following assumption about what is, in any sense, possible points the way to one such strategy of paraphrase: if something might be true, then it might be true in some particular way. It would make no sense to affirm, for example, that there might be life elsewhere in our galaxy, while denying that it is possible that there be life in any particular part of the galaxy, or that there might be life of any particular kind – animal life, or non-animal life – elsewhere in the galaxy. If something is possible, then it is possible that it be realized in a concrete way – perhaps in many alternative ways. The possible-worlds framework begins with this simple assumption, and with the assumption that, in general, statements about what may or might be true can be described in terms of the ways a possibility might be realized. The framework takes alternative specific ways that possibilities might be realized as the primitive elements, out of which propositions – the things that are said to be possible, necessary, or true – are built,

and in terms of which the modal properties of those propositions are defined. The main benefit of this move is that it permits one to paraphrase modal claims in an extensional language that has quantifiers, but no modal auxiliaries, and so in a language in which the semantic structure of the usual modal discourse can be discussed without begging controversial questions about that structure.

I have been arguing for the metaphysical neutrality of the possible-worlds framework: but I should emphasize that I do not mean to suggest that the use of the framework is free of ontological commitment to possibilities (such as ways things might be, counterfactual situations, or possible states of the world). Regimentation clarifies one's commitments, but does not pretend to eliminate them. Furthermore, it must be conceded that the moves made in this regimentation of modal discourse (particularly the move that paraphrases "– might have been true in various particular ways" as "there are various particular ways that – might have been true") are not completely innocent (see Chapter 31, MODALITY, §3.3). As Quine would be the first to emphasize, no strategy of regimentation is neutral in any absolute sense: "The quest of a simplest, clearest overall pattern of canonical notation is not to be distinguished from a quest of ultimate categories, a limning of the most general traits of reality" (Quine, 1960, p. 161). But it is a desideratum of any such project that it be able to accommodate and articulate a range of alternative responses to the questions and puzzles that motivated the project. I think the possible-worlds framework satisfies this desideratum, but the real test of this claim is not in some general methodological argument, but in the fruits of the work that is done with its help.

3 What Are the Semantic Values of Names?

The possible-worlds framework provides the resources to state and clarify both metaphysical and semantic theses. In both cases, I want to argue, the principal conceptual benefit of the apparatus is that it provides an account of a subject-matter that is independent of languages used to describe that subject-matter. Of course, whether the subject is geology or modal metaphysics, we never get away from language – it is just too hard to say very much without using it. But just as we want to distinguish rocks from words (even if we have to use words for rocks to do it), so it is useful to distinguish possibilities from the words used to describe them. To make this distinction is not to beg any questions against the philosophical thesis that the source of all necessity is in language; a conceptual distinction does not foreclose the possibility that one of the things distinguished may in the end be reduced to the other.

To see that possibilities are part of the subject-matter of semantics as well as of modal metaphysics one need only make the following assumptions: First, a central function of an assertion is to convey information, and information is conveyed by distinguishing between possibilities. Second, a principal goal of semantics is to explain how the expressions used to perform speech-acts, such as assertion, are used to convey information – to distinguish between possibilities – and how the way complex sentences distinguish between possibilities is a function of the semantic values of their parts. To understand what is said, for example, in an utterance of "The first dog born at sea was a basset hound," one needs to know what the world would have to be like in order for what was said in that utterance to be true.

These simple assumptions about the goal of semantics might be expressed in terms of truth-conditions: semantics is concerned, among other things, with the truth-conditions of statements, and the way their truth-conditions are a function of the semantic values of their

parts (where semantic values are whatever they must be in order to contribute appropriately to truth-conditions). What are truth-conditions? If we are looking for an answer to this question that identifies a non-linguistic object that semantics can associate with statements, it seems natural to say that the truth-conditions of a statement are the possibilities that, if realized, would make the statement true. We want a conceptual distinction between truth-conditions and any particular forms of expression in which those truth-conditions might be stated, simply because it is useful to theorize about a language in a different language, and when we do so, we want to be able to talk not just about the inter-linguistic relations between the language we are theorizing about and the language we are theorizing in, but about the relation between the language we are theorizing about and the world.

The task of descriptive semantics, in this framework, is to say what kinds of things the semantic values of expressions of various categories are, and to explain how the truth-conditions of sentences (or sentences in context) are a function of the semantic values of their constituents. So to give the semantic value of a proper name is to say what contribution a proper name makes to the truth-conditions of the sentences containing it, where the truth-conditions of a sentence, or a sentence in context, are represented by the set of possibilities that, if realized, would make what the sentence says in that context true. The two answers that Kripke compares – the Millian and the Fregean answers – are made precise in the following ways:

(1) The semantic value of a name is simply its referent; the proposition expressed by a simple sentence containing a name is the proposition that is true in a possible world if and only if that referent has the attribute expressed by the predicate of the sentence.
(2) The semantic value of a name is its sense, which is a concept that applies to, at most, one individual in each possible world (the kind of concept that might be expressed by a definite description). The proposition expressed by a simple sentence containing a name is the proposition that is true in a possible world if and only if the individual to which the concept expressed by the name applies in that world has the attribute expressed by the predicate of the sentence.

Thus far, I have talked only about a question and a framework for clarifying alternative answers to it, and not about arguments in support of one or the other of the answers. The framework is neutral on the question of which of these alternatives, if either, gives a correct account of the semantics of the expressions in English and other natural languages that we identify as proper names. This seems to be an empirical question of no particular philosophical interest, a question that philosophical analysis and argument are not relevant to answering. The way in which the alternative answers are articulated in the framework does point the way to some of the empirical considerations that may be relevant to settling the issue, by making clear what the consequences of those alternatives are; but it appears that no philosophical – certainly no metaphysical – issue hangs on which answer is right. Even though Kripke defended, on empirical grounds, the Millian answer, he nowhere suggested that things had to be this way. For all that he argues to the contrary, we might perfectly well have spoken a language with names that all referred only by having senses that determined referents. It just happens that we do not.

What needed philosophical defense was not the empirical adequacy of the Millian answer, but its coherence. While Kripke did not suggest that philosophical argument could

establish that the answer he favored was correct, he had to answer philosophical arguments that purported to establish that it was incorrect. John Searle, for example, argued that

> the view that there could be a class of logically proper names, i.e. expressions whose very meaning is the object to which they are used to refer, is false. It isn't that there just do not happen to be any such expressions: there could not be any such expressions.… The view that proper names are 'unmeaning marks,' that they have 'denotation' but not 'connotation,' must be at a fundamental level wrong. (Searle, 1969, p. 93)

Michael Dummett makes a similar claim: "there cannot be a proper name whose whole sense consists in its having a certain object as referent, without the sense determining that object as referent in some particular way" (Dummett, 1973, p. 232). These claims are puzzling, in the light of Kripke's way of posing the problem of the descriptive semantics of proper names. It appears that he showed, simply in setting up the alternatives, how to give a coherent specification of a language containing "expressions whose very meaning is the object to which they are used to refer." What kind of argument could show not only that we do not in fact speak such a language, but that we could not possibly do so? To address this question, we need to turn to the second of the two kinds of issues in the philosophy of language that Kripke is concerned with.

4 How Do Names Get Their Semantic Values?

Why do Searle and Dummett think that we could not speak a language of the kind that Kripke described, in which the semantic value of a name is simply its referent? Searle's reason was that "if the utterance of the expressions communicated no descriptive content, then there could be no way of establishing a connection between the expression and the object," no way to answer the question, "What makes *this* expression refer to *that* object?" (Searle, 1969, p. 93). Dummett's reason is similar: "an object cannot be recognized as the referent of a proper name … unless it has first been singled out in some definite way" (Dummett, 1973, p. 232). In both cases, the reason for rejecting the possibility of a certain descriptive semantic thesis appeals to considerations that relate to the foundational question, which asks what it is about the capacities, customs, practices, or mental states of a speaker or community of speakers that makes it the case that an expression has the semantic value that it has. What seems to be suggested is that the hypothesis that a language has a Millian semantics poses a foundational question that cannot be given a satisfactory answer.

But this is not the way either Searle or Dummett put their claims, since they do not separate the two questions, "what is the semantics for names (or the semantic value of a particular name) in the language we speak?" and "what makes it the case that the language we speak (or a particular name in that language) has this semantics?" Once the two questions are separated, it is difficult to see what could rule out the *possibility* that we speak any language that has a well-defined semantics. If a Millian semantics for names can be articulated, why can't a community of speakers adopt the convention to speak such a language?

The assumption implicit in the rejection of the possibility of a Millian semantics is that the two questions we have separated should receive a single answer. Something like a Fregean sense should explain why a name has the particular referent it has, where this is interpreted to mean that it should explain both what it is about the capacities and attitudes

of the speaker that give the name the referent it has, and also what it is that the speaker communicates or conveys in using the name. Kripke charged Frege with conflating these two questions:

> Frege should be criticized for using the term 'sense' in two senses. For he takes the sense of a designator to be its meaning and he also takes it to be the way its reference is determined. Identifying the two, he supposes that both are given by definite descriptions. (Kripke, 1980, p. 59)

Whether Frege is responsible for making this mistake is a question of textual interpretation that I will not comment on, but it is clear, I think, that Searle is concerned with both kinds of questions, and that he takes himself to be following Frege. Searle describes his axiom of identification, a principle that is supposed to be constitutive of singular definite reference, as "a generalization of Frege's dictum that every referring expression must have a sense" (Searle, 1969, p. 80). The principle is about what must be communicated or conveyed (or at least "appealed to" or "invoked") in the utterance of the referring expression; but it is also an attempt to say what it is about the capacities of speakers that explains why their referring expressions have the referents they have. "What I am trying to get at," he says, "is how noises identify objects" (Searle, 1969, p. 83).

If we are implicitly looking for a semantic account of names that answers both questions at once, then the Millian theory that says that the semantic value of a name is simply its referent looks like a non-answer; it seems to be denying the obvious fact that there must be something about the capacities, behavior, or mental state of the users of the name that makes it the case that the name has the referent that it has. On the other hand, the conflation of the two questions masks the fact that the sense theory, interpreted as an answer to the question of descriptive semantics, is also a non-answer to the foundational question. Suppose we were to accept the Fregean thesis that names have the referent that they have because they have a sense that determines a function whose value (at the actual world) is that referent. This simply raises the question: What is it about the capacities, behavior, or mental state of the users of the name that makes it the case that the name has the sense that it in fact has? Whether one is a Fregean or not, the two questions need to be distinguished; and once they are, the way is opened for answers to each that are less easily seen as possible answers to the other: the Millian answer to the descriptive question, and the causal account of reference that Kripke defends as an answer to the foundational question. This latter thesis – that what makes it the case that a name has a certain individual as its referent is that the individual plays a certain role in the causal or historical explanation of the speaker's use of the name – makes no sense as an attempt to specify a semantic value, a candidate to be the meaning or sense of a name, and so it can be taken seriously only after the two questions are distinguished.

Kripke and other defenders of a causal theory of reference were criticized for the vagueness of their thesis. Causal connections are ubiquitous, and it is obvious that there are a great many individuals that are causally implicated in the speaker's use of the name, but that are not by any stretch of the imagination plausible candidates to be the referent. A proper causal theory of reference would have to specify just what sort of causal connection is necessary and sufficient for reference, and that is a notoriously difficult demand. Kripke himself emphasized that he was presenting not a reductive analysis of reference, but only an alternative picture. To some skeptics, this sounded like an evasion. The suspicion was that any sufficiently specific and precise version of the causal theory would be subject

to as many counter-examples as the description theory, and so that the plausibility of the positive alternative to the description theory rested on its lack of specificity. But I think this line of criticism misses the point. What is essential to the alternative picture was the separation of the questions, and the distinction between two different ways in which the extension of an expression might depend on the facts: first, what semantic value an expression has depends on the facts; second, if the semantic value is a sense, the extension of an expression with a given semantic value may depend on the facts. The philosophical work was done in making the distinctions that removed the obstacles to accepting the naïve answers to the questions that were distinguished. If we ask what one has to know to understand a name, the naïve answer is that one must know who or what it names – nothing more. (In contrast, no one would be tempted to give this answer to the analogous question, of what one must know to understand a definite description.) And if we ask how a name comes to name what it names – what, for example, makes "Shakespeare" as we use it a name of the particular person Shakespeare – I think most people would be inclined to point to a historical narrative: his family was called "Shakespeare," or something like that, at the time, and knowledge of him, his plays, and his name were passed down through the generations to us. This is not a particularly exciting philosophical theory, but it doesn't seem wrong, and it does seem incompatible with the kind of answer implied by a description theory. Kripke's causal-chain story is just an articulation of the naive answer, and one that does not add a lot of constructive detail to it. But by separating and clarifying the questions to which these naive answers are answers, he brought out why the theoretical reasons for resisting those answers are bad reasons.

The diagnosis of equivocation is rarely the end of the matter in a philosophical argument. I will sketch a line of argument for the impossibility of a Millian semantics that recognizes the distinction between the two kinds of questions. I don't think this line of argument is successful, but it is instructive to see where it takes the debate.

If there is a credible defense of the thesis that a Millian language is impossible, I think it must challenge the assumption that any well-defined semantics might be the semantics of a language that is used by a speaker, or realized in a speech community. Here is one way that the assumption might be challenged. First, the following seems a reasonable general constraint on the correctness of a claim that a certain semantics is the semantics for the language spoken in a certain community of speakers: if the semantics is correct, then speakers must know, at least for the most part, what, according to the semantics, they are saying. A notion of saying might allow that in some cases one succeeds in saying things using words one does not understand; but it is hard to deny, first, that if one doesn't know what one is saying, then one does not mean what one says, and second, that according to a correct account of what people say in a given speech community, speakers generally mean what they say (not in the sense that they are sincere – believe what they say – but just in the sense that what they say coincides with what they mean).

Second, we may note that it is possible to give a determinate specification of a semantic value without knowing what that value is, even without anyone knowing what the value is. Consider this example discussed by Gareth Evans: let "Julius" be a (rigid) proper name for the person (whoever he or she was) who invented the zip (Evans, 1982, p. 31). Then (assuming that some particular single individual invented the zip) a sentence such as "Julius was born in Minsk" expresses a determinate proposition about a particular individual, but we don't know who the individual is, so we don't know what proposition it is that is expressed. We understand a description of the proposition, and we understand, and may believe, metalinguistic statements about that proposition, such as 'What is said by "Julius was born

in Minsk" is probably false.' We can have beliefs and make assertions about the truth or falsity of whatever proposition is expressed, but (according to this line of argument), we do not thereby assert or believe that Julius was or was not born in Minsk, and we cannot do so unless we know who invented the zip.

Now suppose that one could make a case that our mental relations to particular individuals are in all relevant respects like our relation to Julius – that since we can know individuals only by description, only as whoever or whatever it is that is presented to us in a certain way, we don't ever know, in the relevant sense, who or what it is that we are referring to with the names we use. Then it would seem to follow that, although we can define a language with a Millian semantics, we could never speak one, since we could not have the knowledge required to know what the sentences of such a language say.

I think the first premise of this argument should be conceded: a semantics for the language spoken by a community of speakers cannot be right if it implies that speakers generally do not know what they are saying. It should also be conceded that, according to the Millian semantics for names, as contrasted with the Fregean semantics, speakers do not know what they are saying when they use a name if they do not know who the referent of their name is. But what is it to know what one is referring to? At this point the battleground shifts from semantics and the philosophy of language to the philosophy of mind, where variations on some of the same battles are fought.

Underlying the contrasting answers to the foundational question about reference (What makes it the case that a name has the referent it has?) are contrasting strategies for answering parallel foundational questions about mental states: What makes it the case that a thought – a judgment or an intention – has the content it has, or is about what it is about? The argument sketched above against the possibility of a Millian semantics for an actual language rests on the assumption that thoughts can be about particular things only by expressing general concepts that apply to those individuals. If this were right, then the kind of causal-chain or historical-explanation story that Kripke and Keith Donnellan told to answer the foundational question about semantics would be an answer that detached the determination of the semantic values of expressions from the mental states and capacities of the users of the expressions, and so would be an answer that was vulnerable to this argument. But why should we think that thoughts, any more than names, can be about individuals only by expressing general concepts?

In defense of his principle of identification and his argument against the Millian theory, Searle asks:

> What is it to *mean* or *intend* a particular object to the exclusion of all others? Some facts incline us to think that it is a movement of the soul – but can I intend just one particular object independent of any description or other form of identification I could make of it? And if so, what makes my intention an intention directed at just *that* object and not at some other? Clearly the notion of what it is to intend to refer to a particular object forces us back on the notion of identification by description. (Searle, 1969, p. 87)

The suggestion implicit in these rhetorical questions is that an intention to refer to a particular individual must be explained as a behavioral capacity, the capacity to give a general description, or otherwise identify an individual who is, by fitting the description or being the object identified, the object meant or intended. The only alternative, it is implied, is some kind of obscurantist intentional magic, some kind of movement of the soul.

Even if Searle were right in his suggestion that intentions and other intentional states directed at particular individuals must be explained in terms of the capacities of the agent to identify the individual, this would still not give him the additional premise needed for the argument against the possibility of a Millian semantics for names. For Searle is not arguing that we cannot have intentions and other attitudes toward particular individuals; he is only arguing for a condition that is necessary for having such intentions and attitudes. What he needs for the argument is a constraint on the content of the attitudes we can have; but what he offers instead is a constraint on the conditions under which one can have attitudes with a certain kind of content. Whatever it is that constitutes intending and having knowledge and beliefs about a particular object, "to the exclusion of all others," so long as it is possible to have such intentions, knowledge, and beliefs, it will be possible to understand and speak a language with a Millian semantics.

But in any case, there is no real argument for the conclusion that mental magic is the only alternative to an explanation of intentionality in terms of an agent's capacities to identify. A causal account of intentional content – an explanation that looks back to how mental states came to be, rather than only forward to what those states dispose the agent to do – is equally compatible with a non-obscurantist account of intentionality. A causal account of intentions and beliefs seems, in fact, to be presupposed by the defense of a causal account of reference given by Kripke and Donnellan, since it is argued that speakers not only can refer, but can intend to refer, to particular individuals without being able to describe or identify those individuals. Causal and non-causal accounts of how names get their reference can share the assumption that reference is determined by intentions. The causal theory of reference is causal because it assumes a causal account of the content of the intentions that determine reference.

An argument of Michael Dummett's for the impossibility of a Millian semantics, like the argument I have sketched, bases this conclusion on the impossibility of a certain kind of knowledge: what Dummett calls *bare knowledge of reference*. Here is his characterization: "A *bare* knowledge of the reference of the name *a* will consist … in knowing, of some object, that *a* refers to it, where this is a *complete* characterization of this particular piece of knowledge" (Dummett, 1991, p. 127). I am not sure what Dummett means by a "particular piece of knowledge," or what it is for a characterization of such a piece to be complete; but if we interpret Dummett's arguments in the context of the possible-worlds conception of content, I think it is reasonable to identify his notion of bare knowledge of reference with knowledge of a singular proposition – the proposition that is true in a possible world if and only if a certain particular individual is the referent of the name *a*. An essential step in Dummett's argument for the impossibility of bare knowledge of reference is a claim that is essentially equivalent to Searle's principle of identification: we cannot have what Dummett calls knowledge-what – knowledge *of* a certain individual that it has some property F (for example, knowledge of a certain individual that it is the referent of the name *a*) – unless we have the capacity to describe or identify the object. More strongly, for any true knowledge-what ascription, there must be a true propositional-knowledge ascription whose content is a non-singular proposition that makes the method of identification explicit, and that entails the knowledge-what ascription: a propositional-knowledge ascription on which the knowledge-what ascription "*rests*" (Dummett,1991, p. 130). Now, I am not persuaded that this claim is correct; but even if this much is granted, I don't think it gives one reason to reject the possibility of knowledge of singular propositions. Suppose we grant that one cannot know of some particular individual x that it is F unless for some G one identifies x as the G,

and knows that the G is F. Further, suppose we grant that in a particular case the claim that y knows of x that it is F rests on, and is entailed by, the claim that y knows that the G is F. What has been granted is a claim that certain conditions are necessary, and others sufficient, for having knowledge of a certain kind; but nothing follows from this about the content of that kind of knowledge. If "bare knowledge" of some object, that *a* refers to it, is taken to mean knowledge that can exist in isolation, without knowing anything else about what *a* refers to, then we can grant that bare knowledge of reference is impossible; but that does not imply that knowledge of x, that *a* refers to it, is not knowledge of a particular proposition, a singular proposition, or that it is not possible to have knowledge of such propositions.

We can agree with Dummett that it is a difficult problem to say just what conditions must be met for one to know who or what the referent of a name *a* is, and that the problem is not solved simply by saying that to have such knowledge-what is to know a singular proposition of the form "x is the referent of *a*." But the problem is not that saying this would be wrong; it is just that specifying the content of a knowledge ascription is not the same as saying what it is for that knowledge ascription to be true.

The distinctions, on the level of both speech and thought, between questions about what content is and about how content comes to be determined, help to open up a place in conceptual space for a causal account of reference and of intentionality generally, and provide a rebuttal to arguments for the impossibility of a Millian semantics for a realized language. They do not, of course, end the debate. Once theoretical obstacles to such accounts are removed, examples and untheoretical considerations make a strong *prima facie* case for the claim that some such account is correct; but, as Gareth Evans reminded Kripke, "the deliverances of untutored linguistic intuition may have to be corrected in the light of considerations of theory" (Evans, 1982, p. 76). I think more theoretical considerations also support causal accounts of intentionality, but that is a different part of the story. I want to turn back now to questions about the relation between reference and metaphysical necessity.

5 Names and Essences

Whatever the fate of the debate between them, we have a stark contrast between two pictures of the way we are related, both by speech and by thought, to particular things in the world. On one picture, we can think and talk directly about particular things in virtue of our causal interaction with those things; on the other, our mental and linguistic acts relate us to particular things only by our grasping and expressing purely qualitative concepts that may be instantiated by particular things. The question now is, do these pictures of mental and linguistic representation either presuppose or support some particular conception of the nature of the particular things that we talk and think about, or that instantiate our concepts? Specifically, does the conception of reference that Kripke argues for presuppose or support the particular brand of essentialism that he defends? The two kinds of issues, I will argue, are independent. The only role of the theory of reference in Kripke's arguments for metaphysical conclusions is to help diagnose and rebut fallacious arguments that rest on a conflation of the two kinds of issues.

One of Searle's arguments against the Millian account of proper names was that it (or at least an "uncritical acceptance" of it) leads us into some "metaphysical traps" (Searle, 1969, p. 163). It is suggested that this conception of proper names presupposes "a basic metaphysical

distinction between objects and properties or aspects of objects" (p. 164). Actually, Searle's attitude toward the relation between the metaphysical and the semantic issues is not entirely clear. Does he locate the mistake in the premise – the Millian semantics for names – or in the inference from this premise to a metaphysical conclusion? On the one hand, we are warned against "the original sin of metaphysics, the attempt to read real or alleged features of language into the world" and "the metaphysical mistake of deriving ontological conclusions from linguistic theses" (p. 164); but on the other hand, the fact that the Millian account of names seems to presuppose this metaphysical distinction is part of an argument against that semantic account. It cannot be a good argument against a semantic account of proper names that someone has illegitimately drawn metaphysical conclusions from that account. So perhaps the view is that the Millian theory of names is already a covertly metaphysical thesis: that false metaphysical conclusions are validly drawn from it because the thesis is the result, rather than the occasion, for reading alleged features of language into the world. But if so, can't we separate the semantic from the metaphysical aspects of the thesis, and evaluate the semantic thesis independently of the metaphysical conclusions that are illegitimately drawn from it? If we couldn't make such a separation, then it would not be so clear that it was illegitimate to draw ontological conclusions from linguistic theses.

Searle's target is the Wittgenstein of the *Tractatus*, and not Kripke (whose lectures were given after Searle's book was published). But Kripke does make the kind of metaphysical distinction between objects and properties that Searle rejects; and, of course, he also defends the account of names that Searle argues is the illegitimate source of the metaphysical distinction. I will sketch the way Kripke makes the metaphysical distinction, and then argue that his metaphysical theses are compatible with the Fregean picture of mental and linguistic representation, and so do not presuppose the Millian semantics or the causal theory of reference. What is presupposed in the defense of Kripke's metaphysical conception is only that the two accounts of reference and intentionality not be conflated. I will conclude by considering whether there is a dependence that goes in the other direction: whether the semantic picture that Kripke defends presupposes his metaphysical theses about the relation between individuals and their properties. Here the issues are harder to disentangle, but I will argue that the Millian theory of names and the causal theory of reference are compatible with alternative metaphysical conceptions of individuals and their properties. There is no derivation of metaphysical conclusions from semantic premises.

What is it to make a basic distinction between objects and properties? Searle derided the metaphysical picture of an object as "a combination of its propertyless self and its properties," as well as the contrasting picture of an object as nothing but "a heap or collection of properties" (1969, p. 164); but what do these pictures really come to? Kripke also scorned the same two contrasting pictures, rejecting the assumption that objects are some kind of "bare particulars" or "propertyless substrata underlying the qualities," as well as the claim that they are nothing but bundles of qualities, "whatever that may mean" (Kripke, 1980, p. 52). The possible-worlds framework suggests a way to express the idea that a particular is conceptually separable from its properties without relying on the rejected picture of a bare particular. Properly understood, the issue concerns the modal properties of an individual. Intuitively, it seems clear that ordinary things might have had different properties from the ones they in fact have: Shakespeare might not have written plays, which, in the possible-worlds paraphrase, is to say that there is a possible world in which Shakespeare did not write plays. This possible world is one in which the very person who actually wrote the plays we know and love exists, and did not write plays. So (assuming the modal claim is

right) the property of writing plays is not essential to a particular person who has this property. The same goes for lots of other ordinary properties ascribed to ordinary persons and things; but not to all of them. Shakespeare obviously could not have been someone other than Shakespeare (although he could have been called something else), and according to Kripke he could not have been a member of a different species, or even have had different parents. Now these simple modal claims say nothing about names or reference; but in stating them I am using the proper name "Shakespeare," so the content of what I am saying about counterfactual possibilities might be thought to depend on the semantics for names. If "Shakespeare" were an abbreviation for a definite description, as Russell argued, then the statement that Shakespeare might not have written plays, and its paraphrase, that there is a possible world in which Shakespeare did not write plays, would both be ambiguous. If, for example, "Shakespeare" were an abbreviation for "the most famous Elizabethan playwright," then on one interpretation the claim that Shakespeare might not have written plays would be the claim that there is a possible world in which the person who is the most famous Elizabethan playwright in that world did not write plays. This, of course, is not the claim that Kripke intended to make, and it does not seem, intuitively, that the words used to make the claim are open to such an interpretation. The modal claims, in either their ordinary form or in their possible-worlds paraphrases, do not seem ambiguous, but that is a linguistic intuition which is separable from the modal intuition about the person who is the referent, by whatever means, of the name "Shakespeare," the claim that is expressed by the other reading of the claim that the most famous Elizabethan playwright might not have written plays. Now on what I have been calling the Fregean conception of mental representation, perhaps we cannot have such modal intuitions about particular individuals, since perhaps, on that conception, the only way we can have any thoughts at all about an individual is to have beliefs about whomever it is that is presented to us in some particular way. But this does not matter to the issue, since the metaphysical intuition can be expressed in a perfectly general way: whoever the person is who fits our Shakespeare concept, that person might have been someone who didn't write plays. So the modal theses stand or fall independently of the success or failure of the defense of theses either about the semantics for proper names, or about the way our thoughts relate us to the individuals that are the referents of those names.

If both the Fregean conception of reference and intentionality that Searle favors and the alternative semantic conceptions that he criticizes are compatible with the metaphysical distinction between particulars and their properties, what is the source of his objections to this distinction? I think they derive from a conflation of the two semantic conceptions. By equivocating between the two semantic theses, one can argue from semantic assumptions to a metaphysical conclusion: On the Fregean picture, there is an analytic, and so necessary, connection between the name "Aristotle" and a cluster of properties, and so it is legitimate to conclude, as Searle does, that "it is a necessary truth that Aristotle has the logical sum [inclusive disjunction] of the properties commonly attributed to him" (Searle, 1969, p. 173). This claim, by itself, implies nothing about what must be true of the person Aristotle, and so raises no problems about the traditional distinction between objects and their properties. On the unequivocal Fregean picture, "Aristotle" means, roughly, whoever satisfies the cluster; so according to this semantic hypothesis, if the person Aristotle hadn't satisfied the cluster, he would not have been Aristotle, but he might still have existed. It is only when one combines the Fregean premise with the assumption incompatible with it, that the name "Aristotle" is a Millian name, that one can take the next step in the argument, the inference

from the claim that Aristotle satisfies the cluster, to the conclusion that it is true *of* Aristotle that he satisfies the cluster. As Searle says, "I wish to argue that though no single one of them is analytically true of Aristotle, their disjunction is" (p. 169). It is only when one has the conclusion that the person Aristotle is necessarily connected with the properties used to identify him that one has reason for skepticism about substance, and for the claim that "it is misleading, if not downright false, to construe the facts which one must possess in order to refer as always facts *about* the object referred to, for that suggests that they are facts about some *independently* identified object" (p. 93).

To make a positive case for his modal theses about individuals, all Kripke does is to develop the framework in which the theses can be formulated clearly and separated from theses about names and reference. The rest of the work is done simply by pointing out what seems from an intuitive point of view obviously true, once it is clear what the alternatives are. If Kripke's rhetorical style had been a little different, he might have made this point by saying that he was just assembling reminders, not putting forward theses. What philosophy does, he might have said, is simply to put everything before us (Wittgenstein, 1953, §§126–128).

Not all of the metaphysical claims about individuals that Kripke defends on intuitive grounds are equally compelling. On the one hand, it seems hard to deny that we can make intuitive sense of questions about the potentialities of particular individuals independently of the means used to refer to them. To suppose that Shakespeare never wrote plays is to envision a counterfactual situation in which Shakespeare – the man himself – wrote no plays. Other theses Kripke defends are more controversial from an intuitive point of view, particularly theses that deny that certain things are possible, or equivalently, that affirm that particular things have certain essential properties. Can I coherently suppose that Shakespeare – the man himself – had different parents from the ones he in fact had, or that he was born in a different century? Kripke would argue that if we think we can suppose these things, we are confused: if we think clearly about what we are trying to suppose we will see that these are not coherent counterfactual possibilities. Not everyone will share these intuitions, even after setting aside the bad reasons for resisting them that Kripke points out. The possible-worlds framework does not settle such metaphysical questions, or even tell us how they should be settled; its job is to raise them, and to make clearer what the alternative answers say.

I have argued that Kripke's metaphysical theses do not presuppose his theses on reference and intentionality. What about the other direction? Does the Millian semantics for proper names or the causal account of reference and intentionality presuppose the metaphysical picture that Kripke defends? The theses about names can easily be separated from the more specific essentialist theses, but the general metaphysical issues about the identification of individuals across possible worlds are more difficult to disentangle from the thesis that names are rigid designators whose reference is established by causal interaction between the speaker and the referent. A rigid designator is one which denotes the same individual in all possible worlds; doesn't this presuppose that the same individuals can be found in different possible worlds?

Consider the following anti-essentialist metaphysical picture, a version of the "bundle of qualities" conception of an individual rejected by both Kripke and Searle. According to this conception, a particular individual is just the co-instantiation of a certain set of qualities. If individuals are identified across possible worlds at all, it is only in virtue of some counterpart relation which is definable in terms of the relations between the bundles of qualities

co-instantiated in the different worlds. Consider first the pure Leibnizian version of this metaphysical picture, according to which particular individuals have all of their properties, including their relational properties, essentially. On this conception, not only is it a mistake to think that we can coherently suppose that Shakespeare – the person himself – had different parents, we cannot even make sense of a counterfactual possibility in which he ate a slightly different breakfast than he in fact ate on a certain morning, or even lived in a world in which slightly different events took place years after his death. The thesis that names are rigid designators is perfectly compatible with this uncompromising metaphysics, but the combination of metaphysical and semantic theses has no plausibility. It gives us the conclusion, for example, that the proposition expressed by the statement that Shakespeare wrote plays is one that is true only in the actual world, and so is one that entails every true proposition. Only God could know that Shakespeare wrote plays. We more limited creatures can know that "Shakespeare wrote plays" expresses a true proposition, and we can know that, whatever proposition it expresses, it is necessarily equivalent to the proposition expressed by "Elvis Presley played the guitar"; but so long as we are ignorant of any fact, we cannot know which proposition it is that these sentences express.

Giving up the Millian theory of names would not resolve the problem, since there is no plausibility in the assumption that however we refer to him, Shakespeare – the man himself – could not possibly have failed to write plays, or to eat what he in fact ate for breakfast. A less uncompromising version of this metaphysical picture gives a different account of what it is for an individual to have a property essentially. According to the liberal Leibnizian (Lewis, 1986, ch. 4), to say that Shakespeare – the person himself – might not have written plays is to say that a counterpart of that person in some possible world did not write plays, where the counterpart relation is reducible to some kind of qualitative similarity. The counterpart variation of the Leibnizian metaphysics of individuals is not a thesis about names, but is about the modal properties of individuals, however they are referred to. The difference between the unreconstructed Leibnizian theory and the counterpart version might still be construed as a semantic rather than a metaphysical difference, but it is a difference in the way complex predicates involving modality are to be interpreted, and by itself says nothing about how names are to be understood. But the absurd consequences drawn from the combination of the Leibnizian metaphysics and the Millian account of names were about the propositions expressed by simple sentences involving non-modal predicates. If "Shakespeare" is a Millian name, then "Shakespeare wrote plays" is true in a possible world only if Shakespeare himself wrote plays in that world, which means only in the actual world. The fact that we construe "could have written no plays" in such a way that "Shakespeare could have written no plays" is also true in (and only in) the actual world is beside the point.

One might try to reconcile the Leibnizian metaphysics in its counterpart variation with a version of the Millian semantics for names by reinterpreting the concept of a rigid designator in a way that parallels the reinterpretation of modal predication. Suppose we say that a designator is *quasi-rigid* (relative to a possible world w) if its referent in any possible world w′ is a counterpart of some particular individual in w. Then if "Shakespeare" is a quasi-rigid designator, relative to the actual world, "Shakespeare wrote plays" might be true in some other possible worlds – worlds in which a counterpart of Shakespeare wrote plays. The thesis that names are quasi-rigid designators is a kind of compromise between the Millian theory and the sense theory that accounts for some of the phenomena Kripke brought to our attention, but there are problems with it. First, if (as David Lewis has argued) an individual may have more than one counterpart in the same world, then the semantic value of

a name will not be determined by the individual. Suppose Shakespeare has two counterparts in some possible world, only one of whom wrote plays. Is "Shakespeare wrote plays" true in that possible world or not? To answer this we need to know which of two (or more) quasi-rigid concepts is expressed by "Shakespeare." Second, even if an individual has at most one counterpart in any possible world, a designator might be quasi-rigid relative to one possible world and not relative to another, since two counterparts of the same thing (in different possible worlds) need not be counterparts of each other.

Despite these problems, the basic Kripkean picture of the way the reference of names is determined might still be reconciled with this metaphysical theory. What is essential to Kripke's picture, I think, is the idea that the content of speech acts and mental attitudes may be determined as a function of particular things (and kinds) with which the speakers and thinkers interact. Whatever one's metaphysical presuppositions, it will be agreed that the way content and reference are determined by the facts will be context-dependent, and influenced by general beliefs, purposes, and assumptions. The counterpart-theorist's story about the background against which reference to a particular (world-bound) individual can determine a proposition will be different from the story Kripke might tell, but there is nothing in the counterpart-theorist's metaphysics that prevents him telling such a story.

The dialectic of this last discussion shows that metaphysical and semantic issues cannot be kept completely apart for at least two reasons. First, if semantic and metaphysical theses together yield implausible consequences, it may be a matter of dispute where the source of the problem lies. Second, some metaphysical theories may force a reformulation of claims about semantics and intentionality. But the possible-worlds framework helps to clarify the metaphysical alternatives, and to separate metaphysical from semantic issues so that each can be evaluated on its own terms.

References

Dummett, M. 1973. *Frege: Philosophy of Language*. London: Duckworth.
Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.
Evans, G. 1982. *The Varieties of Reference*. Oxford and New York: Oxford University Press.
Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Salmon, N. 1981. *Reference and Essence*. Princeton, NJ: Princeton University Press.
Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge University Press.
Wittgenstein, L. 1953. *Philosophical Investigations*. New York: Macmillan.

36

Names and Rigid Designation

JASON STANLEY

The fact that natural-language proper names are rigid designators is an empirical discovery about natural language. However, unlike other empirical discoveries about language made in the past few decades, it is one which has been taken to have great philosophical significance. One reason for this is that it has helped simplify the formal semantical representation of ordinary modal discourse. But the central reason is that the discovery threatens a certain picture, the descriptive picture, of the content of names, upon which a great deal of philosophy was premised. (See Chapter 35, REFERENCE AND NECESSITY.)

This chapter is mainly intended to be a survey of both the background and contemporary discussion of this discovery. However, the survey takes place in the context of an evaluation of the extent to which the discovery that English proper names are rigid itself threatens the descriptive picture of the content of names. The goal is to show that the exact philosophical significance of the discovery that natural-language proper names are rigid designators is still, and should still be, a matter of controversy.

§1 discusses different explications of rigidity. §2 is devoted to a sketch of the development of the notion of rigidity. §3 is a discussion of the descriptive picture of the content of names. In §4, Kripke's argument for the thesis that natural-language proper names are rigid is outlined, as well as an argument based upon this thesis against the descriptive picture. Finally, the remaining three sections cover various possible defenses of the descriptive picture.

1 Rigidity

Rigidity is a semantic property of an expression. More specifically, it has to do with the evaluation of that expression with respect to other possible situations (or 'worlds'). There are many subtle issues involved in the notion of evaluating an expression with respect to a possible situation, some of which we will discuss in this chapter.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

But there are also some simple confusions about this notion. Before we begin our discussion of rigidity, it is important to dispel one such confusion.

On one way of understanding evaluation of a sentence with respect to another possible world, a sentence is true with respect to another possible world just in case, if the sentence were uttered in that other possible world, it would be true. However, this is decidedly *not* how to understand the notion of evaluation with respect to another possible world which underlies our modal discourse.

The correct notion of evaluation of a sentence with respect to another possible world involves considering the sentence as uttered in the *actual world*, rather than as uttered in other possible worlds. When the sentence is uttered in the actual world, it expresses some semantic value which is determined by how the words are used by speakers in the actual world. This semantic value is then evaluated with respect to other possible worlds. What the nature of the entity is which is evaluated with respect to other possible worlds – whether it is a "proposition" (what is said by an utterance of the sentence) or some other entity – is a difficult question, and one which we will address at the end of this chapter. But for now it is only important to note, as a preliminary to our discussion of rigidity, that what is at issue in evaluating a sentence with respect to another possible world does not involve considering that sentence as uttered in that other possible world, but rather considering the sentence as uttered in the actual world.

How an expression e is used by speakers in other possible situations is thus irrelevant to the question of what the extension of e is when evaluated with respect to those other possible situations. For instance, what the denotation of "Cayuga Lake" is with respect to another possible world has nothing to do with how the speakers of that world – if there are any – use the expression "Cayuga Lake." It just has to do with which object Cayuga Lake is in that world. Now that this possible confusion has been eliminated, we may turn to the notion of rigidity.

According to Kripke's characterization of rigidity, "a designator d of an object x is rigid, if it designates x with respect to all possible worlds where x exists, and never designates an object other than x with respect to any possible world."[1] This characterization, as Kripke intends, is neutral on the issue of the extension of the designator d in possible worlds in which x does not exist. That is, if d is a designator which satisfies the above criteria, there are three possibilities left open for d's extension in worlds in which x does not exist. First, d could designate nothing with respect to such possible worlds. Second, d could designate x in all such possible worlds (despite x's non-existence in those possible worlds). Third, d could designate x with respect to some such worlds, and designate nothing with respect to other such worlds.

These three possibilities determine three different species of rigidity. However, only the first two species deserve discussion; a designator in the third class is a hybrid, and there is no reason to countenance such expressions. In the rest of our discussion, I will not consider designators in this third class left open by Kripke's characterization of rigidity.

The first species of rigidity, corresponding to the first of the above possibilities, includes all and only those designators d of an object x, which designate x in all worlds in which x exists, and designate nothing in worlds in which x does not exist. Following Nathan Salmon (1982, p. 4), let us call these *persistently rigid designators*.

The second species of rigidity, corresponding to the second of the above possibilities, includes all and only those designators d of an object x, which designate x in all worlds in which x exists, and designate x in all worlds in which x does not exist; or, more simply,

designate x with respect to every possible world. Again following Salmon, let us call these *obstinately rigid designators* (Salmon, 1982, p. 4).

There are expressions which are uncontroversially rigid in both of the above senses. For instance, consider Kripke's class of *strongly rigid designators* in *Naming and Necessity* (Kripke, 1980, p. 48). This class contains the rigid designators of necessary existents. That is, this class contains all and only those designators d of an object x which exists in all possible worlds, which designate the same thing in all possible worlds (that is, x). For example, the descriptive phrase "the result of adding two and three" is a strongly rigid designator, since its actual denotation, namely the number five, exists in all possible worlds, and the phrase denotes that number with respect to all possible worlds. Strongly rigid designators clearly belong to both of the above classes.[2]

At several points in this chapter considerations in support of the notion of obstinate rigid designation over that of persistent rigid designation will be advanced. However, it is unclear to what degree issues about persistent rigidity versus obstinate rigidity are substantive, rather than merely disguised terminological discussions about how best to use the expression "evaluation with respect to a world." There is a sense of this expression in which it seems to presuppose the existence of the denotation in the world; and if someone is using the expression in this sense, then persistent rigidity might be the more appropriate notion. If, on the other hand, one has a purely semantical understanding of "denotation with respect to a world," then the fact that the semantic rules directly assign a denotation to an expression might lead us to think that even in worlds in which that object does not exist, it is still the denotation of the relevant expression. But these are certainly just terminological issues.[3]

A further distinction is often made in discussions of rigidity: that of Kripke between *de jure* rigidity and *de facto* rigidity (see Kripke, 1980, p. 21, n. 21). An expression is a *de jure* rigid designator of an object just in case the semantical rules of the language unmediately link it to that object. All other rigid designators of objects are *de facto* rigid designators of them. To give an example from Kripke, the description "the smallest prime" is supposed to be *de facto* rigid, because it is not metaphysically possible for there to be a smallest prime distinct from the actual smallest prime, that is, two. The fact that "the smallest prime" denotes the same object in every world flows not from semantics, but from the metaphysical fact that mathematical facts are true in all metaphysically possible worlds. If, on the other hand, the semantical rule for a term t takes the form of a stipulation that it denotes a certain object x, then t is *de jure* rigid, since it is part of the semantical rules that it denotes that object.

The intuitive content of *de jure* designation lies in the metaphor of "unmediated" reference. A rigid *de jure* designator is supposed to denote what it denotes without mediation by some concept or description. A *de facto* rigid designator, on the other hand, is supposed to denote what it denotes in virtue of its denotation meeting some condition. That is, a *de facto* rigid designator denotes via mediation of some concept or description.[4]

The core notion of rigidity has been taken by philosophers to be *de jure*, obstinate rigidity. This is the notion which lies at the center not only of Kripke's work, but also of David Kaplan's work on direct reference (see Kripke, 1980, p. 21, n. 21; Kaplan, 1989a; and, most explicitly, Kaplan, 1989b, pp. 569–571; see also Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS). We will give some (albeit not so weighty) reasons in future sections for preferring obstinate rigidity over persistent rigidity. But we shall see already in the next section why the *de jure* character of rigidity is thought to be important. For rigidity arose in the development of the semantics of Quantified Modal Logic (henceforth QML), and in

particular, as a part of the explanation of the proper treatment of variables in QML. In that context, there is no question that *de jure* rigidity is the relevant concept.

2 Rigid Designation and Quantified Modal Logic

The pre-theoretic notion of rigidity began its life as a concept in the semantics for QML. In particular, rigidity arose in connection with the 'objectual' interpretation of QML, where the quantifiers were taken to range over objects, rather than non-constant functions. Even more specifically, rigidity was relevant to issues concerning Quine's "modal paradoxes," raised as objections to the coherence of QML. In this section, I will attempt to show where the notion of rigidity enters into the attempt to give a coherent and natural semantic interpretation to QML.

One of the first issues which arose in QML was what the proper intended interpretation of quantification should be. The two camps in the 1940s were the conceptual interpretation, championed by Alonzo Church and Rudolf Carnap, and the objectual interpretation, championed by Ruth Barcan Marcus.[5] But while Church, Carnap, and Barcan Marcus and others were developing axiom systems for QML, Willard Van Orman Quine was busy attempting to demonstrate their incoherence.

Quine raised two influential objections to QML (Quine, 1943).[6] According to the first of these objections, quantification into modal contexts violated fundamental logical laws. According to the second (and obviously related) objection, if QML and its intended interpretation could be so formulated as to evade the first objection, then it would inexorably carry with it unpalatable metaphysical commitments.[7] Since the defenders of QML partially defined their own positions against the first of these objections, something must be briefly said about it here. Following this we will outline the conceptual interpretation of QML, and then the objectual interpretation, explaining how their original espousers evaded Quine's worry.[8]

According to the principle of substitution, for any terms a and b, if "a = b" is true, then for any formula φ containing "a," the result of replacing one or more occurrences of "a" by "b" does not change the truth-value of φ.[9] However, according to Quine, QML essentially involved a violation of this principle. For "nine = the number of planets" is true. Furthermore, "Necessarily, nine = nine" is true. But the result of substituting "the number of planets" for the first occurrence of "nine" in "Necessarily, nine = nine" yields a falsity, namely, "Necessarily, the number of planets = nine."

Quine took the failure of substitution in modal contexts also to demonstrate the failure of existential generalization in QML. That is, Quine took the failure of substitution to show that the inference from "$\Box Fa$" to "$\exists x \Box Fx$" is illegitimate. The reason Quine thought that a failure of substitution demonstrated the failure of existential generalization is that he thought that substitutability by co-referential terms was a *criterion* for the legitimacy of quantifying in.[10]

Here is one reason why Quine thought that the substitutability of co-referential terms in a linguistic context C was a criterion for the legitimacy of quantification into C. Consider a quotational context, such as:

(1) The first sentence of the (English translation of the) *Duino Elegies* is "Who, if I cried out, would hear me among the angels' hierarchies?"

Inside such a quotational context, substitution of co-referential terms fails to preserve truth-value. For example, (1) is true, but (2), which results from (1) by the substitution of co-referential terms, is false:

(2) The first sentence of the (English translation of the) *Duino Elegies* is "Who, if Rilke cried out, would hear Rilke among the angels' hierarchies?"

Thus, substitution of co-referential terms fails in quotational contexts.

But it is also illegitimate, according to Quine, to quantify into such contexts. To see this, consider the sentence:

(3) There is something x such that "Who, if x cried out, would hear x among the angels' hierarchies?" is the first sentence of the *Duino Elegies*.

Sentence (3) is false. The reason (3) is false is, as Quine is fond of pointing out, that the quoted sentence in (3) names not some sentence which results from replacing 'x' by a term, but rather a sentence containing the symbol 'x.' That is, a quotation such as "'x flies'" denotes the result of concatenating the symbol 'x' with the word 'flies,' not the concatenation of some replacement term for 'x' with "flies." Thus, for Quine, it is illegitimate to quantify into quotational contexts.[11]

But Quine does not simply conclude from the failure of both substitution and quantifying into quotational contexts that substitution is a criterion for quantifying in. For Quine, the failure of substitution in a linguistic context demonstrates a deep incoherence in quantifying into such contexts. For in giving the semantics of a quantified sentence, one must avail oneself of the notion of satisfaction; the sentence is true just in case some object satisfies the relevant open sentence. Yet, for Quine, the failure of substitution shows that there is no available notion of satisfaction in terms of which one can define the truth of such sentences. There is no notion of objectual satisfaction for quantifying into quotational contexts, for instance, because such contexts are sensitive not just to objects, but also to how they are named.

Thus, for Quine, the failure of substitution in modal contexts demonstrated that there was no appropriate notion of objectual satisfaction for open formulas such as "$\Box$Fx." For the failure of substitution seemed to show that whether or not an object satisfied an open, modalized formula depended upon how the object was named. Quine hence thought there was a similarity between modal and quotational contexts: in both cases, what matters is how the object is named, rather than just the object itself. Quine concluded that there was no way of giving a coherent semantics for sentences such as "$\exists x \Box Fx$," since there was no available notion of satisfaction in terms of which one could define the truth of the sentence. He hence declared that quantification into modal contexts was illegitimate (since incoherent), and that existential generalization fails.

There is also a historical reason for Quine's analogy between modal and quotational contexts: for Quine's target, Carnap, wished to explicate necessity in terms of the analyticity of certain sentences. That is, Carnap in *Meaning and Necessity* believed that to say that a certain proposition was necessary was "really" to say, of a certain sentence, that it was analytic (Carnap, 1988, p. 174). Thus, according to Carnap, a construction such as (a) "really" expressed (b):

(a) Necessarily, bachelors are unmarried men.
(b) "Bachelors are unmarried men" is analytic.

So, according to Carnap, modal contexts were really disguised quotational contexts. If so, then quantifying into modal contexts seems tantamount to quantifying into quotational contexts.

There are several responses which have been given to Quine's challenge. One response stems from the interpretation of QML which emerged from the work of Church and Carnap. According to this approach, variables in modal languages ranged over individual concepts, describable (in contemporary terms) as functions (possibly non-constant) from possible worlds to extensions. The principle of substitution, on this approach, was interpreted as licensing not substitution of terms for two extensionally equivalent individual concepts (that is, functions which yield the same denotation in the actual world), but rather, substitution of terms which denote the same individual concept.

Now, "nine" and "the number of planets" do not express the same individual concept, for though they are extensionally equivalent, there are possible situations in which the extension of "the number of planets" is different from the extension of "nine." Thus, the principle of substitution does not license the substitution of "the number of planets" for "nine," on this account of QML. Furthermore, any two expressions which do express the same individual concept (are "L-equivalent," in Carnap's terms) will be substitutible, even in modal contexts.

This 'conceptual' interpretation of QML thus has a systematic, logically consistent account of the notion of the satisfaction of an open-modal formula (cf. Church, 1943; Carnap, 1988, §§43 and 44). On the conceptual interpretation of QML, one can take the quantifier in "$\exists x \Box Fx$" to range over individual concepts. In this case, the relevant notion of satisfaction is satisfaction by individual concepts, rather than objects.[12]

However, the conceptual interpretation of QML does not seem to accord with our natural interpretation of QML. The sentence:

(4) $\exists n$ ($\Box$n numbers the planets)

is intuitively false on a natural reading of the quantifier (cf. Garson, 1984, pp. 265–267). The reason it seems false to us is that, according to a very natural reading of (4), what it asserts is that there is some object which necessarily numbers the planets. However, on the conceptual interpretation, (4) is true, because the individual concept expressed by "the number of planets" will satisfy the open formula:

(5) $\Box$n numbers the planets

since, in every possible world, the number of planets numbers the planets.

What such examples suggest is that the natural reading of quantification into modal contexts is as quantification over objects, rather than over individual concepts. If we wish to capture this intuition, then we should think of, say, an existential quantification into an open-modal formula (henceforth OMF) as true just in case some object satisfies the relevant modal condition.[13] On this account, which we shall call the objectual interpretation of QML, the first-order quantifiers range only over objects, rather than over concepts.

According to the objectual interpretation, a sentence such as "$\exists x \Box Fx$" is true just in case some object is necessarily F. But what about Quine's worry? Can the objectual interpretation supply a natural account of the satisfaction of OMFs?

An OMF, such as "□Fx," is, on the objectual conception, satisfied by an assignment just in case the object which that assignment assigns to 'x' is necessarily F, that is, is F with respect to every possible situation, *irrespective of any names of that object*. We are not to understand the satisfaction of such an OMF "substitutionally," as satisfied by an assignment, just in case, for some name a of the object which that assignment assigns to 'x,' the sentence, "□Fa" is true. Rather, we are to read "□Fx" as satisfied by an assignment s just in case the object which that assignment assigns to 'x' satisfies F with respect to every possible situation.

This understanding of the satisfaction clause for OMFs undercuts Quine's objection to the coherence of quantifying into modal contexts. For Quine's worry can only arise if objectual satisfaction is characterized in terms of the truth of closed sentences containing names of the alleged satisfiers. Only if objectual satisfaction is given such a substitutional construal is it relevant to the coherence of quantifying into modal contexts that two closed modalized sentences, differing only in containing different names for the same object, may differ in truth-value.[14]

If such a notion of an object satisfying a predicate necessarily indeed makes sense, then it is possible to quantify into modal contexts despite the failure of substitution. Of course, Quine's *other* objection to QML is that, where the necessity in question is metaphysical, this notion involves a dubious metaphysic of essentialism. But discussion of this question will take us too far away from the topic of rigidity (see Chapter 34, ESSENTIALISM, and Chapter 35, REFERENCE AND NECESSITY).

This construal of the satisfaction of OMFs, combined with possible-world semantics, naturally brings with it an interpretation of variables according to which they are *de jure* rigid designators. To see why this is so, consider a sentence of QML such as "$\exists x \Box(\text{Exists}(x) \rightarrow \text{Rational}(x))$."[15] According to the objectual interpretation of QML, this sentence is true just in case there is some assignment function which assigns to the variable 'x' an object o which, in every possible situation, satisfies the open formula "$\text{Exists}(x) \rightarrow \text{Rational}(x)$." The evaluation of the truth of the sentence hence involves, relative to an assignment function, evaluating the open formula "$\text{Exists}(x) \rightarrow \text{Rational}(x)$" with respect to every possible situation. Since, in each possible situation, we are considering whether or not the object o satisfies the formula, we need to ensure that the variable 'x' denotes o in all of the possible situations. That is, on the objectual interpretation of QML, when taken with respect to an assignment s, variables are rigid designators of the objects which s assigns to them. The reason that variables are *de jure* rigid designators is because there is nothing else to the semantics of variables besides the stipulation that, when taken with respect to an assignment s which assigns the object o to a variable, it designates o in every possible situation.[16]

If we understand variables as rigid designators (with respect to an assignment), then the following version of substitution is validated:

(6) $\forall x \forall y[x = y \rightarrow [\varphi \leftrightarrow \psi]]$

(where φ differs from ψ only in containing free occurrences of "x" where the latter contains free occurrences of "y"). For even if φ and ψ contain modal operators, the rigidity of the variables will guarantee the inter-substitutability of "x" and "y."

The situation is slightly more complicated in the case of terms. Quine's challenge is to validate, not just (6), but also the fully schematic version of substitution:

(7) $t = s \rightarrow (\varphi \leftrightarrow \psi)$

(where φ differs from ψ at most in containing occurrences of t where the latter contains occurrences of s, and no free variables in t and s become bound when t and s occur inside φ and ψ). But where t and s are replaceable by non-rigid designators, then (7) will, in the modal case, fail to be valid; thus the defender of the objectual interpretation who wishes to preserve full classical substitution must disallow non-rigid terms from her language.

There are also other motivations for restricting the class of terms to rigid ones on the objectual interpretation. For example, to do so would allow a uniform treatment of the class of terms. If all terms are rigid, then non-variables can be treated in the semantics as free variables whose interpretation does not depend on assignments.[17] Another reason is that, if one allowed non-rigid designators, one would have to restrict universal instantiation (UI) to rigid designators to retain (6), and some might hold that such a restricted UI rule is unappealing. Finally, non-rigid terms raise further technical problems which, though certainly solvable, nevertheless complicate the semantics.[18]

At this point the reason for the introduction of terms which directly represent objects is purely technical – it is a technical response to a logico-semantical dilemma. If one wishes to preserve classical substitution, as well as the objectual conception of satisfaction, then one must ensure that one's variables and terms are rigid. In availing ourselves of such terms, there is no commitment to thinking that any terms in ordinary language are rigid. Rigid terms only play the role, at this stage, of desirable formal-semantical tools, which allow us a better grasp of the objectual notion of satisfaction, as well as an explanation of the validity of classical substitution.

However, if we wish QML to serve as a representation of ordinary modal discourse, then the rigidity constraint on terms may seem problematic. Without a philosophical justification of this restriction, or a semantical argument to the effect that natural-language terms are rigid, this restriction is *ad hoc*. If natural-language singular terms are non-rigid, then the extra logico-semantical complexities which attend the addition of non-rigid terms into QML will either have to be accepted as realities or used as a basis for rejecting its coherence.[19]

Even in the late 1940s it was recognized that a philosophical/semantical argument demonstrating the rigidity of natural-language terms would be desirable.[20] However, it was not until the seminal work of Saul Kripke in 1970 that a fully explicit argument for this conclusion was forthcoming. But Kripke's ambitions went far beyond demonstrating that natural-language terms are rigid. For Kripke used the notion of rigidity as a basis for quite substantive claims about the nature of intentionality. It was thus with Kripke that the *philosophical* construal of rigidity began.

3 The Descriptive Picture

According to the picture of intentionality attacked by Kripke, the way our words hooked onto an extra-linguistic reality was via description. That is, a name such as "Aristotle" denoted the person, Aristotle, because the name was associated with a series of descriptions (such as "the last great philosopher of antiquity") which were uniquely satisfied by the person Aristotle. More relevant for our purposes, however, is Kripke's attack on the descriptive picture of the *content* of proper names. According to this, the content of a name was given by the description which fixed its referent. That is, what someone said when they uttered a sentence such as "Aristotle is F" was a descriptive proposition to the effect that, say, the last great philosopher of antiquity, whoever he was, is F.

Kripke (1980) first demonstrated that ordinary-language proper names were rigid. He then used this feature of names as part of a larger attack on a certain version of the above picture of content.

In the next section, we will discuss how Kripke used rigidity to attack the descriptive picture. But before we do so, it is important to gain an understanding of what the descriptive pictures of intentionality and content are. In particular, we will distinguish between two different versions of the descriptive picture which are often not distinguished in the literature.

The problem of linguistic intentionality, in one of its forms, is the question of what it is in virtue of which an expression has the reference it does. According to the first descriptive picture of linguistic intentionality, what it is in virtue of which a primitive expression has the referent it does is that it is associated with a set of descriptions, in purely general, non-indexical, or particular involving terms. These descriptions are uniquely satisfied by an entity which then counts as the reference of that term.

A less problematic and more commonly held version of the description theory dispenses with the requirement that the descriptions which fix referents must be given in purely general terms. According to this version, which is most explicit in the works of Strawson and Dummett, but at least implicit in Frege, the descriptions which fix referents can, and indeed often must, contain non-descriptive elements.[21]

It is worthwhile to mention briefly a motivation for the latter picture of intentionality. One might think that, in the case of demonstrative reference, one has reference without any description. But this is merely a myth. Suppose I point to a brown table, and say, "This is brown." It is not my pointing alone which fixes the reference of the occurrence of "this," for my finger will also be pointing at the edge of the table, or a small brown patch on the table. Rather, a factor in fixing the reference of my demonstrative is that I intend to be demonstrating some object whose identity criteria are those of tables, rather than those of small brown patches or edges. Such identity criteria play a crucial role in overcoming the massive indeterminacy of ostensive definition. It is for their specification that descriptive material is required.[22] But this insight in no way requires that we ignore the non-descriptive element inherent in true demonstrative reference (see Chapter 39, OBJECTS AND CRITERIA OF IDENTITY).[23]

A final relevant factor which distinguishes descriptive accounts of intentionality from each other has to do with the role of the social. According to Russell, as well as the account of descriptive intentionality attacked by Kripke (1980), a term refers, in the mouth of a speaker, to that object which satisfies the descriptions the *speaker* associates with the term. However, according to other traditional descriptive accounts, such as that of Strawson (1959, pp. 151 ff.), what is relevant is not which descriptions the speaker associates with the term, but rather, which descriptions are associated with the term in the language community. On this latter, more plausible account, a use of a term in the mouth of a speaker refers to the object it does in virtue of her participation in a language community which associates certain descriptions with that term that are uniquely satisfied by the object in question.[24]

There are thus two different versions of the descriptive picture, one according to which the descriptions must be in general terms alone, and another in which they may contain irreducible occurrences of demonstrative and indexical expressions. Each of these two versions has two sub-versions: one according to which it is the descriptions the speaker associates with a term which are relevant for determining the reference of terms she uses, and the other according to which it is the descriptions the language community of the speaker associates with the term which determine the reference of the term when she uses it.

Each of these versions corresponds to a theory of the content of sentences containing proper names. On the first picture, utterances of sentences containing proper names express descriptive propositions, where the relevant descriptions only contain expressions for general concepts. According to the second version of the description theory, utterances of sentences containing proper names also express descriptive propositions. However, these descriptive propositions typically are also irreducibly indexical propositions. So, on this latter account, a sentence such as "Bill Clinton is F" would state some proposition equivalent to what is expressed by "The *present* president around here of the United States is F."[25]

If the descriptive picture is true, then, for each expression in our language, we possess, *a priori*, uniquely identifying knowledge about its referent. Such a premise is more than just a useful tool in epistemological and metaphysical theorizing. For if the descriptive picture is true, then we have a rich store of *a priori* knowledge. This makes more plausible a classic picture of philosophy, according to which it proceeds by *a priori* methods. The Kripkean challenge to the descriptive picture is thus not merely a challenge to an empirical thesis, but also threatens to undermine deeply rooted conceptions of the nature of philosophy.

4 Kripke's Argument and the Rigidity Thesis

I will not go into great detail in this chapter about Kripke's larger critique of the descriptive picture of intentionality and content, as the issue is covered in another chapter in this volume (see Chapter 35, REFERENCE AND NECESSITY). In this section I will first describe an argument, due essentially to Kripke, for the thesis that names are rigid designators. I will then conclude with an argument from rigidity against the descriptive picture of content.

One of the central contributions of Kripke (1980) lay in the argument that natural-language proper names are rigid designators (where "rigid designator" is taken in the first, neutral sense of §1). In what follows, we will go through this argument. More exactly, what we will motivate is the following thesis, which I will call RN, the Rigid Name thesis:

(RN) If N designates x, then N designates x rigidly

where "N" is replaceable by names of English-language proper names. Throughout the argument for RN, it will be assumed that variables under assignments are rigid designators, and it will be argued from this assumption that natural-language proper names are also rigid designators.[26]

According to the neutral characterization of rigidity, a designator D of an object x is rigid just in case, for all possible worlds w, if x exists in w, then D designates x in w, and if x does not exist in w, then D does not designate something different from x in w. There are thus three ways in which a designator D of an object x could fail to be rigid:

(a) There could be a world in which x exists, but is not designated by D.
(b) There could be a world in which x exists, but D designates something else.
(c) There could be a world in which x does not exist, and D designates something other than x.

It will be argued that each of these possibilities is ruled out in the case in which D is a proper name.

Before we proceed with the argument, it is worth noting that no separate proof is required for (b). Given that proper names designate, at most, one thing in each world, any situation in which x exists, but D designates something else will be a situation in which D does not designate x. That is, every (b) situation is an (a) situation. Thus, the demonstration that (a) is incompatible with D being a proper name will suffice to show that (b) is incompatible with D being a proper name.

So let us first argue that if "a" is a proper name designating x, then, in any world in which x exists, x is designated by "a." Suppose not, that is, suppose "a" designates x, and (a) is true. Then the following is the case:

(8) $\exists x\,[x = a\ \&\ \Diamond\,(x \text{ exists } \&\ x \neq a)]$

But (8) seems false when "a" is a proper name. Plugging an actual proper name in for "a" in (8) should make this clear:

(9) There is someone who is Aristotle but he could exist without being identical with Aristotle.

This is intuitively false. Thus, it seems that if N is a proper name designating x, then, if x exists in a world, then N designates it. So, we are done with case (a) as well as (b).

Now, let us turn to the argument that if "a" is a proper name designating x, then, in any world in which x does not exist, "a" does not designate something other than x. Suppose not, that is, suppose "a" designates x, and (c) is true. Then the following is the case:

(10) $\exists x\,[x = a\ \&\ \Diamond\,(a \text{ exists and } a \neq x)]$

But (10), like (8), seems false when "a" is a proper name. Substituting an actual proper name for "a" in (10) should make this clear:

(11) There is someone who is Aristotle but Aristotle could exist without being him.

Like (9), (11) also seems intuitively false. Thus, it seems that if N is a proper name designating x, then, if x does not exist in a world, then N does not designate anything else. So we are done with case (c), and the argument for (RN).

The argument for (RN) exploits speaker's intuitions about the truth-value of instances of (8) and (10). In the case of normal proper names, it seems true that, when substituted for "a" in (8) and (10), a false sentence results. (RN) is thus an empirical claim about natural language. As such, it has been challenged. That is, some have maintained that there are true instances of (8) and (10). However, the proper names that are typically considered are somewhat elaborate, involving issues in metaphysics that are beyond the scope of this chapter. The literature on "contingent identity-statements" will thus not be discussed in what follows.[27]

In the above description of Kripke's argument, I have been using the expression "rigid designator" in the sense of a term which denotes its actual denotation in all possible worlds

in which that denotation exists, and nothing else in other worlds. But there are also some considerations which some have felt mitigate in favor of the thesis that names are obstinately rigid designators. For instance, Kripke (1980, p. 78) gives as an example the sentence:

(12) Hitler might never have been born.

Sentence (12) is true. But (12) is true just in case the sentence, "Hitler was never born" is true when evaluated with respect to some possible world. If "Hitler" does not denote anything with respect to that world, then, unless one gives sentences containing non-denoting terms truth-values, it will be impossible to make the sentence "Hitler was never born" true in that world. But, if "Hitler" denotes Hitler in that world, then, despite the non-existence of Hitler in that world (or perhaps because of it), the sentence "Hitler was never born" can be true in that world.

This argument is, however, unimpressive. For it relies on the thesis that sentences containing non-denoting terms receive no truth-value. If one said that sentences containing non-denoting terms were false, then analyzing "Hitler was never born" as the negation of "Hitler was born" in a world in which "Hitler" is non-denoting would yield the correct prediction.[28]

Many have adverted, at this point, to a more indirect argument, one which exploits the analogy between tense and modality. A tense-logical obstinately rigid designator is one which denotes the same thing at all times, regardless of whether or not that thing exists at the time of evaluation. That proper names should be treated as tense-logical obstinate rigid designators is supported by the Montagovian example:

(13) John remembers Nixon.[29]

Example (13) can be true, as uttered in 1995, despite Nixon's non-existence at the time of utterance. Such evidence is taken, by the tight analogy between tense and modality, to support the modal logical obstinancy of proper names.[30]

However, examples such as (13) only demonstrate that proper names can denote individuals existing prior to, but not during, the time of evaluation. If proper names are to be true tense-logical obstinate rigid designators, then proper names of objects which exist subsequent to, but not during, the time of evaluation, should nonetheless denote at the time of evaluation. But this does not seem to be the case. For instance, consider the name "Sally," introduced in 1995 to denote the first child born in the twenty-first century. In the case of such a name, it is dubious that it denotes, as evaluated in 1995. For it is metaphysically likely that the future is open, and not already determined. If so, then there is no fact of the matter, in 1995, as to what the reference of "Sally" is now. Thus, it is unclear whether proper names are tense-logical obstinately rigid designators.

Whatever the outcome of the debate concerning the obstinancy of proper names is, it does seem that proper names are rigid designators. This would suggest that what fixes the referent of a proper name is not a non-rigid description, but rather something else. If so, then the descriptive account of intentionality would seem to be false.

This argument, as Kripke recognized, is, however, too swift. For it collapses once one makes Kripke's useful distinction between a description giving the content of a name and merely fixing its referent. If the description fixes the referent of a name then there is no

commitment to saying that the name denotes an object in other possible worlds in virtue of that object satisfying the description. On this picture, the description fixes the referent, which is then the denotation of the proper name, even in worlds in which the referent does not satisfy the description. Thus, there is no direct argument from rigidity against the descriptive picture of intentionality.

The case differs, however, with the descriptive picture of content. For there does seem to be an argument from rigidity against the thesis that the content of a proper name is descriptive. For suppose that the content of the proper name "a" is descriptive. In particular, suppose that its content is given by the non-rigid description "DD." Then the content of a sentence which results from replacing "N" by "DD" should stay unchanged, since "N" and "DD" have the same content. But, given that "N" is rigid and "DD" is not rigid, (14) and (15) do not have the same content, as (14) is true and (15) is false:

(14) N might not have been DD.
(15) N might not have been N.

Therefore, substitution of "DD" for "N" does not preserve truth-value, and hence also does not preserve content. Hence "DD" and "N" do not, after all, have the same content.

Let us take a concrete example. Suppose that the name "Aristotle" has the same content as the description, "the last great philosopher of antiquity." Then, replacement of "Aristotle" by "the last great philosopher of antiquity" should preserve content. But:

(16) Aristotle might not have been the last great philosopher of antiquity.
(17) Aristotle might not have been Aristotle.

differ in content, since (16) has a true reading (for instance, there is a reading of (16) where it is true because Aristotle might have died as a child, in which case he never would have become a philosopher at all), and (17) has no true reading.[31] Thus, "Aristotle" and "the last great philosopher of antiquity" are not inter-substitutable, and hence do not have the same content.

It thus seems that Kripke's demonstration that proper names are rigid also shows that they do not have descriptive content. An obvious next step is the thesis, which Kripke attributes to John Stuart Mill, that the content of a proper name is simply its denotation. However, Kripke does not, from rigidity alone, conclude that Millianism is correct;[32] rather, he only commits himself to the following minimal thesis, which I shall henceforth call the *Rigidity Thesis*, or RT:

> The rigidity of proper names demonstrates that utterances of sentences containing proper names, and utterances of sentences differing from those sentences only in containing non-rigid descriptions in place of the proper names, differ in content.[33]

If RT is correct, then the descriptive account of content would seem to be false. In the rest of this chapter, I shall focus on various ways of defending the descriptive account of content. In the next section, I will discuss a version of the descriptive account of content which is compatible with RT. After that, I will discuss critiques of RT.

5 The 'Actualized' Description Theory

RT raises a *prima facie* difficulty for descriptive theories of content. Since the most plausible meaning-yielding descriptions seem to be non-rigid, RT seems to demonstrate that descriptive accounts of content are false. However, this appearance is misleading. RT does not demonstrate that all descriptive accounts of content are false. In particular, RT is only incompatible with one of the two descriptive accounts of content distinguished in §3. As we shall see in this section, though RT is indeed incompatible with the thesis that the content of proper names can be given by description in purely qualitative, general terms, it is not incompatible with the more traditional descriptive account of content, according to which the descriptions which give the content of proper names may contain indexical expressions.

RT is incompatible with the purely qualitative description-theory of the content of proper names. For consider plausible meaning-yielding descriptions for an ordinary proper name, such as "Aristotle." Since the meaning of an expression is what one knows in virtue of which one is competent with that expression, such descriptions must be the things that are known by those competent with the expression. Examples of such descriptions are "the last great philosopher of antiquity," or "the teacher of Alexander." But these are non-rigid descriptions. RT is incompatible with such descriptions matching proper names in content.

On the other descriptive account of content considered in §3, the descriptions which give the content of proper names may contain indexical expressions: that is, expressions occurrences of which denote fixed parameters of a context. For instance, "I" denotes the speaker of a context, "now" the time of the context, and "here" the place of the context. (See also Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS.) But once one broadens one's perspective to include modal evaluation, it seems natural to add the word "actual" to the list. That is, once one is in the context of possible-worlds semantics, "actual" indicates the world of the context.[34]

If so – that is, if "actual" is an indexical – it would be bizarre, on an account of content according to which the descriptions which give the content of proper names may contain indexical expressions, to disallow its appearance in the content-yielding descriptive expressions. But descriptions which contain the word "actual" are rigid. That is, a description such as "the actual F" rigidly denotes the object which is in fact F, even in worlds in which that object fails to be the unique F. Indeed, someone sympathetic with this account of the descriptive picture of content, as well as RT, would simply conclude that the descriptions which give the content of proper names must contain the indexical "actual." Furthermore, on this account of content, it would not even be a surprising fact that the relevant descriptions must be "actualized," since, on this account of content, the arguments for the thesis that meaning-yielding descriptions must contain indexical expressions (for instance, Strawson's consideration of symmetrical universes) straightforwardly generalize to the modal case.

Now, if proper names are *de jure* rigid designators, then even this descriptive account of content would be false, for actualized descriptions do not "unmediatedly" designate. That is, a description such as "The actual teacher of Alexander" designates Aristotle via mediation of some concepts.[35] There are several responses to this point.

The first response to this point is that the argument for RN given in §4 does *not* (and was not, by Kripke, intended to) demonstrate that proper names are *de jure* rigid, but merely

that they are rigid. Second, given the metaphorical nature of the notion of mediation, it is difficult to see how one *could* argue for such a conclusion. Finally, there are examples of proper names which do seem, relatively uncontroversially, to rigidly designate what they designate "via mediation."

The first of these points is obvious. The statement of RN does not mention the notion of *de jure* designation. Furthermore, nothing in the argument for it would fail if proper names were only *de facto* rigid.

To grasp the second point, consider the case of indexicals, which are rigid designators. Does the word "I" designate what it designates via mediation, or not? Kaplan (1989a) seems to think it does not.[36] But "I," whenever it is used, designates the agent of the context. Though there are difficulties in making precise the notion of" agenthood" here, it is difficult to see how it could be that "I" designates "unmediately," given the linguistic rule that it is to designate the agent of the context. Perhaps there is a notion of mediation according to which "I" unmediatedly designates. But if so, it needs to be made more precise before an argument for such a conclusion can be evaluated.

Finally, there are examples of proper names which, if the notion of mediation is coherent, do seem to designate mediately what they designate. Consider, for example, the following example, due to Gareth Evans (1985b). Suppose we wish to discuss what the world would have been like if the zip had not been invented. In particular, we wish to discuss what would have happened if the inventor of the zip had died at birth. Not knowing who the inventor of the zip is, we introduce a name "Julius," by the following reference-fixing stipulation:

(S) Reference("Julius") = The inventor of the zip

and then go on to theorize about what would have happened had Julius died at birth, and had failed to invent the zip.

Evans's intuition is that "Julius" is a rigid designator. That is, according to Evans, (18) has no true reading, but (19) does:

(18) Julius might not have been Julius.
(19) Julius might not have been the inventor of the zip.

If so, then "Julius" is an example of a proper name which designates what it does via mediation, and is hence not *de jure* rigid.[37]

Given Evans's example, it seems implausible to maintain that it is a feature of the semantic category of proper names that they are *de jure* rigid. Of course, on the descriptivist account, *no* proper name is *de jure* rigid, which, given the slight oddity of "Julius"-type names, may seem worrisome. Nonetheless, what is important to note for our purposes is that there is no argument from rigidity alone against a traditional descriptive account of the content of proper names. Issues of rigidity are simply independent of the question of whether names have descriptive content.[38]

None of this would be news to Kripke. Kripke never argued that his modal considerations refuted every version of the descriptive account of content. Michael Dummett has, however, leveled more direct challenges to Kripke's conclusions.[39] Though Dummett agrees that Kripke has shown an important difference between English proper names and descriptions, he has challenged Kripke's contention that the difference in question always makes a difference to what is said. In particular, according to Dummett, the rigidity of proper names

does not affect the content of modally "simple" sentences, that is, sentences not containing modal terms. In other words, Dummett challenges the truth of RT.

Dummett's early views on rigidity can be separated into two doctrines. The first, which is a negative doctrine, is that rigidity does not make a difference to the content of simple sentences. The second, which is a positive doctrine, is that the phenomena which the notion of rigidity is intended to capture can be accounted for by a stipulation that terms which Kripke would classify as rigid take an obligatory wide scope with respect to modal operators.

In the next section I will discuss Dummett's positive doctrine, as well as Kripke's decisive objection to it. In the final section I will turn to a more promising line of argument against RT along essentially Dummettian lines.

6 Names and Wide-Scope

Consider again (16) and (17), which were used to show that "Aristotle" and "the last great philosopher of antiquity" have different contents. As we saw in §4, (17) has no true reading, whereas (16) does. If one assumes that "Aristotle" is rigid, whereas "the last great philosopher of antiquity" is not, then one can account for this contrast between the two expressions.

The point of Dummett's positive doctrine is that one can account for the distinction between (16) and (17) without supposing a difference in semantic value between "Aristotle" and "the last great philosopher of antiquity." That is, one can account for the distinction without supposing that "Aristotle" is rigid, whereas "the last great philosopher of antiquity" is not. According to Dummett, all that the distinction between (16) and (17) demonstrates is that there is a *syntactic* constraint on terms such as "Aristotle," which forces them to take wide scope with respect to modal operators.

Here is how Dummett's positive doctrine accounts for the distinction between (16) and (17). (17) and (the true reading of) (16), properly regimented (and abstracting from irrelevant detail), come out, on Dummett's view, as:

(16′) For some x such that Aristotle = x [$\Diamond$ x ≠ the last great philosopher of antiquity]
(17′) For some x, y such that Aristotle = x and Aristotle = y [$\Diamond$ x ≠ y]

(16′) is true because there are possible situations in which the actual denotation of "Aristotle" died as a child; whereas, given the rigidity of variables, (17′) is false. Thus, Dummett's positive doctrine accounts for the distinction between (16) and (17) without postulating a semantic difference between proper names and definite descriptions. Indeed, if Dummett's positive doctrine is correct, proper names can be identified with definite descriptions which take an obligatory wide scope with respect to modal operators.

On Kripke's account, the difference between (16) and (17) is attributed to a difference in the semantic values of the expressions "Aristotle" and "the last great philosopher of antiquity." (17) has no true reading because "Aristotle" is rigid; that is, it is associated with a (perhaps partial) constant function from worlds to objects, whereas (16) does have a true reading, since "the last great philosopher of antiquity" is not rigid; that is, it is associated with a non-constant function from possible worlds to objects. On Dummett's account, no difference in semantic value is required in order to explain the distinction between (16) and (17). It is simply a syntactic feature of proper names that they take wide scope with respect to modal operators; but, in all semantic respects, proper names are like descriptions.

However, Dummett's positive account is problematic, as the following argument by Kripke demonstrates. Suppose "t" is an expression which Kripke would classify as rigid, and "t′" is a non-rigid description which, according to Dummett, has the same content as "t." Consider now the following discourse:

(20) t is t. That's necessary.
(21) t is t′. That's not necessary.

Both (20) and (21) are true, given that "t" is an expression which Kripke would classify as rigid and "t′" is not rigid. But on Dummett's account, it is difficult to see how this could be so.

The central issue in interpreting this discourse is what the content of the occurrence of 'that' is. There are two possibilities. First of all, the occurrences of 'that' might refer to some 'value' of the preceding sentences, either the proposition it expresses ('what it says'), or some other semantic feature. The second possibility is that the occurrence of 'that' refers to the preceding sentences themselves, that is, it is replaceable by a quote-name of the preceding sentences. In each case it is difficult to see how Dummett's account could make both (20) and (21) true.

Suppose the first of these possibilities to be the case. That is, suppose that the occurrence of 'that' in (20) denotes some value of the preceding sentence 't is t.' Then, since both discourses are true, by Leibniz's Law, the value denoted by the occurrence of 'that' in (20) must be different from the value denoted by the occurrence of 'that' in (21), since the values have different properties (one is necessary, while the other is not). But Dummett's positive doctrine gives us no explanation of this fact. According to Dummett, one can explain rigidity facts by a syntactic stipulation that certain terms – those which Kripke classifies as rigid – take an obligatory wide scope with respect to modal operators. But no such operators occur in the initial sentences of (20) and (21). Therefore, Dummett's positive account predicts that there should be no differences in semantic value between these two sentences. But if the two occurrences of 'that' denote some semantic value of the preceding sentences, then the two sentences are associated with different semantic values, *contra* the predictions of Dummett's positive account.

So let us suppose, then, the second of the above possibilities to be the case; that is, suppose that the occurrence of 'that' in (20) denotes the sentence, "t is t." Similarly, suppose that the occurrence of 'that' in (21) denotes the sentence, "t is t′." In this case, we could replace the second sentences in (20) and (21) by:

(22) "t is t" is necessary.
(23) "t is t′" is not necessary.

(22) and (23) are true. But again, on Dummett's positive account, it is not possible to see how this could be the case. For there is no way for any of the occurrences of the term "t" to take wide scope with respect to modal operators, since they all occur within quotation marks.[40]

What Kripke's argument seems to show is that no syntactic account of the distinction between proper names and definite descriptions is possible. Thus, the difference between proper names and definite descriptions must be attributed to a difference in the semantic values they receive. Indeed, one might use this argument of Kripke's to establish RT. For even in the case of unmodalized sentences, replacing a rigid designator by a non-rigid designator

will typically result in a sentence which differs in truth-value in some possible world. One can exploit this to provide an argument for RT.

To see this, consider the following discourses, both true:

(24) Aristotle was not a philosopher. That would be true in a situation in which Aristotle died as a baby.
(25) The last great philosopher of antiquity was not a philosopher. That would not be true in a situation in which Aristotle died as a baby.

Using the same reasoning as in Kripke's argument, it follows that the sentences "Aristotle was not a philosopher" and "The last great philosopher of antiquity was not a philosopher" must have different semantic values. For one value, when evaluated with respect to a situation in which Aristotle died as a baby, is true, while the other, when evaluated with respect to that same situation, is not true. Thus, since the values have different properties – one is true with respect to the world in question, while the other is not true – by Leibniz's Law they must be different.

What this argument of Kripke's establishes is that whatever it is that is evaluated with respect to different metaphysically possible worlds in the case of "Aristotle was not a philosopher" differs from whatever it is that is evaluated with respect to different metaphysically possible worlds in the case of "The last great philosopher of antiquity was not a philosopher." Furthermore, it is clear that similar demonstrations can be given for other cases in which a non-rigid description might seem to have the same content as a rigid designator.

But RT might still seem worrisome. According to RT, sentences containing definite descriptions have different contents from the sentences which result from replacing those definite descriptions by any rigid expressions, even when the sentences are unmodalized. But there are *prima facie* counter-examples to this. For example, the sentences:

(26) The president of the USA came to dinner.
(27) The actual president of the USA came to dinner.

do not, on the face of it, seem to say different things; rather, the difference between utterances of (26) and (27) seems to lie in their pragmatic force. Yet "the president of the USA" is non-rigid, and "the actual president of the USA" is rigid.

Furthermore, (26) and (27) pose a problem for Kripke's argument in this section. For (26) and (27) differ in truth-value with respect to some metaphysically possible worlds. Utterance (26) is true in a world in which George Bush came to dinner, Bill Clinton did not, and George Bush won the 1992 election. Utterance (27) is not true with respect to such a situation. But it seems over-hasty to conclude from this that utterances of (26) and (27) say different things.

Examples such as (26) and (27) might lead one to the view that the semantic differences between rigid and non-rigid expressions do not imply that they must differ in content, as well as to the thesis that the differences in modal semantic value – that is, whatever is evaluated in other possible worlds – do not necessarily lead to a difference in content. Yet these reactions presuppose a distinction between semantic value and content which requires greater explication before it can be developed into a serious response to RT. It is to this task which we now turn.

7 Assertoric Content and Ingredient Sense

In this section, I will introduce and motivate Dummett's distinction between assertoric content and ingredient sense. I will then use this distinction in briefly suggesting a line of critique against RT.

The *assertoric content* of an utterance of a sentence is what is said by that utterance; it is also the object of belief, doubt, and other propositional attitudes. (See Chapter 14, PROPOSITIONAL ATTITUDES.) Assertoric contents are the fundamental bearers of truth-value. They are not true or false relative to a time or a place. Mary's belief that the sun is shining is not true at some times, false at others. What Mary says when she says that the sun is shining is not true in America, false in Australia. It is true or false, as Frege says, *tertium non datur*.

The *ingredient sense* of a sentence is what that sentence contributes to more complex sentences of which it is a part. The ingredient sense of a sentence is thus that sentence's compositional semantic value. It is the semantic value we must assign to a sentence in order to predict correctly the conditions under which more complex constructions in which it occurs are true. As Dummett notes, ingredient sense is what formal semantic theories are concerned to explain.[41]

Once one makes the distinction between ingredient sense and assertoric content, the possibility arises that the ingredient sense of a sentence might differ from its assertoric content. There are several ways in which this possibility might be realized. First of all, it could be the case that sentences which have the same assertoric content nonetheless contribute different things to more complex sentences containing them.[42] That is, it could be the case that sentences with the same assertoric contents have different ingredient senses. Second, it could be the case that the ingredient sense of a sentence cannot serve as its assertoric content, because it is not the sort of object which is fit to be believed or asserted. As we shall soon see, both of these situations in fact obtain.

Consider, first, the former of these possibilities, that is, that two sentences which have the same assertoric content differ in ingredient sense. Each of (29)–(31) has the same assertoric content as (28):

(28) The president is Bill Clinton.
(29) The current president is Bill Clinton.
(30) The president here is Bill Clinton.
(31) The actual president is Bill Clinton.

The difference between each of (29)–(31) and (28) is not truth-conditional, but pragmatic. In each of (29)–(31), a presupposition is present which is not present in (28).

But these presuppositions are cancelable. The sentences can be true, even if the presuppositions fail. Indeed, in any context c, an utterance of each of (29)–(31) is true in c just in case an utterance of (28) is true in c. On a natural construal of the expression "truth-condition," each of (29)–(31) has the same truth-conditions as (28), and hence has the same assertoric content as (28).

However, as the following sentence-pairs demonstrate, the two sentences in each of the sentence-pairs have different *ingredient senses*:

(32) It will always be the case that the current president is Bill Clinton.
(33) It will always be the case that the president is Bill Clinton.

(34) Everywhere, it is the case that the president here is Bill Clinton.
(35) Everywhere, it is the case that the president is Bill Clinton.
(36) Necessarily, it is the case that the actual president is Bill Clinton.
(37) Necessarily, it is the case that the president is Bill Clinton.

In each of these sentence-pairs, the first sentence is true, but the second false. Thus, each of (29)–(31) contributes different things to more complex sentences of which they are a part than (28). But then, given that utterances of them have the same assertoric content as utterances of (28), we have shown that utterances of two sentences can have the same assertoric content, while nonetheless differing in ingredient sense.

Consider now the second of these possibilities, namely that ingredient senses are not the sort of objects which can be identified with assertoric contents, things believed and asserted. As the following examples show, and as Lewis (1981) points out, this too is the case:

(38) It will be the case that the sun is shining.
(39) Somewhere, the sun is shining.
(40) In the future, there might be a miracle somewhere.

In each case, what the embedded sentence contributes to the interpretation of the whole sentence is not something which could plausibly be identified with an assertoric content, something fit to be believed or asserted. In the case of (38), the embedded sentence "the sun is shining" must express a function from times to truth-values. In the case of (39), the embedded sentence must express a function from places to truth-values. Finally, in the case of (40), the embedded sentence must express a function from world, time, and place triples to truth-values.

But, as we have seen, functions from times or places to truth-values are not fit to be things believed or asserted. Mary's belief that the sun is shining does not vary in truth-value from one time to another, or from one place to another. It is true or false, *tertium non datur*. Therefore, ingredient senses are not fit to be assertoric contents.[43]

Let us now sum up our conclusions so far in this section. First, we have seen that sentences can have the same assertoric content, while differing in ingredient sense. Second, we have seen that ingredient senses are not the sort of objects which can be regarded as assertoric contents. Keeping these facts in mind, let us now turn to how these facts bear on RT.

In the original argument for RT, we inferred, from the fact that "Aristotle is Aristotle" and "Aristotle is the last great philosopher of antiquity" embed differently in modal contexts – that is, (16) and (17) differ in truth-value – that the two sentences have different contents, and hence that "Aristotle" and "the last great philosopher of antiquity" have different contents. Yet, once the assertoric-content/ingredient-sense distinction is made, it is clear that this sort of inference is invalid. From (16) and (17), it is only legitimate to infer that "Aristotle is Aristotle" and "Aristotle is the last great philosopher of antiquity" have differing ingredient senses. Similarly, it is only legitimate to infer from (16) and (17) that "Aristotle" and "the last great philosopher of antiquity" have different semantic values. But as we have seen, this does not demonstrate that replacement of one with the other typically yields a sentence with a different assertoric content. That is, such facts as (16) and (17) do not demonstrate the truth of RT.

But what about the Kripkean argument for RT given in the last section? In the case of (20) and (21), and (24) and (25), the initial sentences were not embedded in modal

contexts. Nonetheless, the Kripkean argument established that the sentence containing the rigid designator, and the sentence resulting from it by replacing the rigid designator by a non-rigid designator, corresponded to different values. However, the Kripkean argument only demonstrates RT if the values in question are assertoric contents, or propositions, rather than ingredient senses. For, as we have seen, it is perfectly possible for two sentences to differ in ingredient sense, yet for utterances of them to have the same assertoric content.

The fundamental question in evaluating the Kripkean argument is what the denotations of the occurrences of 'that' are in the relevant discourses. If such occurrences of 'that' denote the assertoric content of the preceding sentences, then the argument does indeed demonstrate RT. If, however, such occurrences of 'that' denote the ingredient senses of the preceding sentences, then the argument only demonstrates that the preceding sentences differ in ingredient sense, a fact perfectly consistent with their coinciding in assertoric content.

Now, there is no question that such occurrences of 'that' can denote the assertoric content of the occurrences of the preceding sentences. This is precisely what the denotation of 'that' is in such contexts as:

(41) The sun is shining. That's asserted by John.
(42) The sun is shining. That's believed by Mary.

Our question is thus: do all such uses of 'that' denote the assertoric contents of the occurrences of the preceding sentences, or do they sometimes denote the ingredient senses of the preceding sentences?

That the latter is the case can be seen from the following two examples:

(43) The sun is shining. That will be true, but it isn't true now.
(44) The sun is shining. That's true somewhere, but it isn't true here.

In order for (43) to be true, the occurrence of 'that' must denote a function from times to truth-values. Similarly, in order for (44) to be true, the occurrence of 'that' must denote a function from places to truth-values. But, as we have seen, such entities are certainly not assertoric contents, things believed and expressed. Such examples hence show that some such occurrences of the word 'that' denote the ingredient senses, rather than the assertoric contents, of the preceding sentences.[44]

The fact that the word 'that' sometimes denotes the ingredient sense, rather than the assertoric content, of the preceding sentence allows the Dummettian to respond to the Kripkean argument as follows. What she would maintain is that the occurrences of 'that' in (20), (21), (24), and (25) denote, not the assertoric content of the preceding sentences, as in the occurrences of 'that' in (41) and (42), but rather the ingredient sense. Since a difference in ingredient sense does not imply a difference in assertoric content, the Kripkean argument fails to demonstrate RT.

For the Dummettian, then, the Kripkean argument fares about as well as the following argument for the thesis that utterances of (28) and (29) always have different assertoric contents:

(45) The current president is Bill Clinton. That will always be true.
(46) The president is Bill Clinton. That won't always be true.

It would be over-hasty to conclude, from this argument, that utterances of (28) and (29) must have different assertoric contents.[45] Similarly, according to the Dummettian, it would be over-hasty to conclude, from (20) and (21) alone, that "t = t" and "t = t′" have different assertoric contents.

What a friend of RT must show is some disanalogy between the argument from (20) and (21) to the conclusion that "t is t" and "t is t′" do not have the same assertoric content, and the argument from (45) and (46) to the conclusion that (28) and (29) have different assertoric contents. There are two ways in which she could proceed. First, she could argue that, in modal contexts, the relevant uses of 'that' do denote the assertoric contents of the preceding sentences. Alternatively, she could argue that, unlike the case of (45) and (46), a difference in the particular ingredient sense, or semantic value, denoted by these occurrences entails a difference in assertoric content.

According to the opponent of RT, the object of modal evaluation, like the object of temporal evaluation, is not a proposition or assertoric content. To make her position clear, she must first provide some clear account of the assertoric-content/ingredient-sense distinction. Then, she must provide an account of assertoric content which distinguishes it in relevant ways from the object of modal evaluation.[46]

Conclusion

As we have seen, given the possibility of actualized descriptions, there is no argument from rigid designation against the description theory of names. The more interesting question, however, is the status of RT. What I have tried (ever so briefly) to motivate is the view that RT is not as innocent as many philosophers believe. The classic Kripkean argument in its favor fails. That is not to say that RT is false: for instance, it may be that the best theory of content entails it.[47] On the other hand, there may be substantive empirical or methodological objections against it. But I am afraid that these are issues which we must leave for future philosophy of language to decide.[48]

Notes

1 This characterization of rigidity is from a letter from Kripke to Kaplan, cited on p. 569 of Kaplan's "Afterthoughts" (1989b).

2 There is another notion of rigidity occasionally suggested in the literature according to which a term is rigid just in case it refers to the same object in all possible worlds in which it refers at all. But this is consistent with the actual denotation of a rigid designator existing in some possible world, yet unnamed by that designator. This possibility is ruled out by Kripke's general characterization of rigidity. In what follows, "rigidity" will instead be used in accordance with Kripke's general characterization.

3 Besides the issues that will be discussed in later sections, there are other issues in philosophical logic which may push one to prefer one or the other characterization of rigidity. For instance, if one defines necessity as truth in every world, then, to capture the intuitive necessity of "Bill Clinton = Bill Clinton," one might wish to allow "Bill Clinton" to denote Bill Clinton with respect to every possible world (in which case one would prefer the characterization of rigidity as obstinate rigidity). Alternatively, one could exploit another notion of necessity, *viz.* non-falsity in every world. This would allow "Bill Clinton = Bill Clinton" to lack a truth-value in some possible worlds without thereby becoming contingent, hence removing the need to treat designators as obstinately rigid to preserve the necessity of "Bill Clinton = Bill Clinton." Similar issues arise with respect to

the characterization of validity. However, here, too, it is difficult to see any substantive issues. As Kripke (1963, p. 66) writes, "For the purposes of modal logic we hold that different answers to [these questions] represent alternative *conventions*. All are tenable."

4 However, eliminating the metaphor of mediation in the characterization of this distinction is no easy task. Furthermore, as will become clear in later sections, it is unclear how the distinction between *de jure* and *de facto* rigid designation generalizes to other expressions.

5 Because of space considerations, I will not discuss the latter's use of substitutional quantification in explicating quantification into modal context.

6 For discussions of Quine's objections, see Fine (1989) and Kaplan (1986). See also Richard (1987). There is a substantial body of contemporary literature on this topic.

7 So perhaps it is not really correct to call these two different objections.

8 I am here, as below, *not* using "objectual" in the sense of the distinction between objectual and *substitutional* quantification, but rather in the sense of the soon-to-be-explicated distinction between quantification over individuals versus quantification over concepts.

9 Here, "a" and "b" and "φ" are being used as schematic letters replaceable by metalinguistic names for object language expressions, and "' is being used for quasi-quotation. I will use "' as normal quotation and quasi-quotation, leaving it to context to disambiguate. In general. I will be lax about use/mention.

10 Kaplan (1986) calls this "Quine's Theorem." See pp. 231–238 for a reconstruction of Quine's (1943) arguments, and Kaplan's critique of it. See also Fine (1989).

11 It is illegitimate *simpliciter* to quantify into contexts in which the quotation is ordinary English quotation. However, Kaplan (1986) introduces a new quotation device, which he called "arc quotes," and showed how to make sense of quantification into them (see §§7 ff.).

12 Furthermore, on the conceptual interpretation of QML, there are ways to rescue substitution of co-extensional expressions in extensional contexts, and even to rescue a quantified version of extensional substitution of the form:

(*) $\forall x \forall y\,(x = y \rightarrow (\varphi \leftrightarrow \psi))$

(where φ differs from ψ in containing free occurrences of 'x' where ψ contains free occurrences of 'y', and φ and ψ are extensional). According to Carnap, for example, both terms and variables are systematically ambiguous. To each term, there corresponds both an extension and an intension (something which yields, at every possible world, an extension). In addition, to each variable, there correspond both *value extensions* and *value intensions.* The value intensions of a variable are the set of intensions of expressions which are admissible substitution instances of that variable, and the value extensions are the set of extensions of expressions which are admissible substitution instances of that variable (see Carnap, 1988, §10). Since in extensional contexts all that is relevant are the value-extensions of variables and the extensions of terms, once the notion of "extensional context" has been appropriately inductively defined (say, as a wff of non-modal first- or higher-order calculus), both the fully schematic version of extensional substitution and the version of substitution containing quantifiers can be preserved. This is Carnap's "Method of extension and intension" (1988, ch. 1). Church avoids having to give expressions and variables a double interpretation, choosing instead to follow Frege in relativizing their interpretations to contexts. For a discussion of these matters, see Fine (1989, pp. 267 ff.). For an old attack on the *metaphysical* coherence of the conceptual interpretation, see Quine (1947).

13 For simplicity's sake, I am speaking here only of non-vacuous existential quantifications into modal formulas with one free variable (so it is appropriate to speak of truth and falsity, rather than satisfaction). I will occasionally make such simplifying assumptions without comment.

14 Of course, such a primitive relational sense of necessity is analogous to Quine's (1956) primitive relational sense of propositional attitude verbs. Quine himself later noticed (1977) that his reconstruction of quantification into propositional attitude contexts could be used in this way to defend the coherence of quantification into modal contexts.

15 Here, "exists" is a primitive predicate which is true of an object with respect to a possible situation just in case that object is in the domain of the possible situation.

16 Missing this point, Quine (1977, p. 8) asserted that the notion of rigidity by itself presupposes the notion of an essential property: "A rigid designator differs from others in that it picks out its object by essential traits." A careful reading of Kripke's discussions of transworld identification (e.g., 1980, p. 44) might have dispelled his belief in this.

17 This is for non-complex terms. If the language contains rigid complex terms – rigid descriptions – the interpretation of terms which are not variables may, of course, depend upon an assignment function.

18 For instance, even the free-logical rule of universal instantiation can fail for languages with non-rigid terms (cf. Garson, 1984, pp. 262–263). Furthermore, the introduction of non-rigid terms complicates completeness proofs for systems of QML, since standard completeness proofs rely on substitution facts (cf. Garson, 1984, pp. 287–289).

19 In retrospect, the latter option seems only to be motivated if one accepts Quine's rather curious idolatry of classical quantification theory. There are many ways of restricting classical substitution to account for non-rigid terms, either by restricting substitution to atomic formulas, or by reformulating quantification theory in terms of complex predicates and restricting substitution to complex predications (for this latter option, see Robert Stalnaker, 1977; 1995). See also, for a development of the appropriate proof theory for a language with complex predicates and non-rigid designators, Fitting (1993, §3; 1991).

20 For instance, Arthur Smullyan (1947; 1948) argued, against Quine's logico-semantical objection to QML, that once one recognizes that descriptions are to be treated on Russellian lines, rather than as terms, then Quine's objection fails. Smullyan is thus the first person explicitly to suggest that natural-language terms are such that classical substitution holds for them. However, since Smullyan wrote years before Kripke's development of the semantics of QML, he cannot be credited with the discovery that natural-language names are rigid, since he did not possess the resources to define the notion of rigidity. Furthermore, he provided no argument to the effect that natural-language terms are rigid. A similar point holds for Barcan Marcus. Though she derived (quantified) versions of the necessity of identity (Barcan 1947; see esp. theorem 2.32), she did not, at that time, have the notion of rigidity, since she had neither an explicit semantics in mind, nor any sort of philosophical or semantic argument about natural language. She does suggest (1961) that natural-language names are mere "tags" for objects, but she neither provides the sort of arguments required for the establishment of this thesis, nor possesses the semantical apparatus necessary to characterize the notion of rigidity. Nonetheless, the work of Smullyan, Barcan Marcus, and also Frederick Fitch certainly provided much of the necessary impetus for the later development of these notions. For an excellent discussion of their role in the history of the notion of rigidity, see Scott Soames (1995).

21 For instance, for Strawson, descriptive identification is based upon demonstrative identification:

> [The supposition that where the particular to be identified cannot be directly located, its identification must rest ultimately on description in general terms] is false. For even though the particular in question cannot itself be demonstratively identified, it may be identified by a description which relates it uniquely to another particular which can be demonstratively identified. (1959, p. 21)

Nonetheless, for Strawson's anti-skeptical arguments to succeed, he must be assuming that successful reference requires uniquely identifying knowledge given by description. Dummett directly challenges the thesis that for Frege, the sense of each proper name can be given by a description (see, to cite one example, the appendix to ch. 5 of his 1981a). For Frege, his belief that a change in reference entails a change in sense demonstrates that he did not ascribe to the "description in general terms alone" account.

22 The case is more difficult in the case of "I" and "here," for their reference is "guaranteed." See Evans (1982, chs 6 and 7) for an attempt to fit an account of these words into a model more closely paralleling perceptual demonstratives than seems, *prima facie*, to be possible, and see Lucy O'Brien (1995) for a recent critique of Evans's account.

23 Though see Kripke (1980, p. 58) for a challenge to this paragraph.

24 This distinction between different descriptive pictures of intentionality is relevant for Kripke's epistemological critique of descriptive theories of intentionality, though not his argument from rigidity.

25 Of course, to bring this fully in line with the description theory, we would also have to analyze the place name "the United States."

26 Furthermore, I will use "possibilist" quantifiers (that is, quantifiers whose domains are not restricted to worlds, but rather range over all actual and possible objects) as well as a primitive existence predicate ("x exists") which is true of an object at a world just in case that object is in the domain of that world (i.e., exists in that world).

27 See, for example, Allan Gibbard's discussion (1975) of "Goliath" and "Lumpl."

28 To rescue the necessity of identity, one would be forced to reformulate some clause in the semantics. One method is to replace the identity axiom schema by its free-logical counterpart. Alternatively, one could redefine the necessity operator, as in, for instance, van Benthem (1983, ch. 12, pp. 136–137), to restrict evaluation of the embedded sentence to worlds at which there exist the denotations of constants in the sentence, and values of the free variables of the sentence.

29 See Montague (1974, p. 126). This example, too, is not fully convincing, since "remembers" is intensional. A slightly better example is "Aristotle is currently the most-read philosopher."

30 Cf. Salmon (1982, pp. 37–39) for a longer discussion. Evans (1985a) has challenged the analogy between tense and modality.

31 That is, (17) has no true reading where the possibility in question is *metaphysical* possibility. Throughout, all occurrences of modal expressions should be read as expressing metaphysical possibility.

32 Kripke's argument that the content of a proper name is only its denotation depends more on the epistemological arguments he gives in Lecture II (1980). We will not discuss these arguments here.

33 This is only a rough statement of the actual thesis. For one may have a coarse-grained account of content, where, say, logical contradictions say the same thing. In this case, utterances of sentences which express logical contradictions, such as "John is tall and it is not the case that John is tall," would say the same thing as utterances of sentences with the name replaced by a non-rigid designator. But this is obviously not an objection to Kripke, for if such an account of content is endorsed, then the statement of the rigidity thesis would have to be modified to capture more adequately Kripke's intention.

34 Formally, the logic of the sentential operator "actually," which is the modal logic analogue of the temporal indexical "now," has been much investigated. Classic papers in this area include Segerberg (1973), Davies and Humberstone (1980), and Hodes (1984a; 1984b). For recent books on the subject, see Graeme Forbes's excellent (1989), which uses rigidifying operators such as "actually" to dispense with quantification over (and hence ontological commitment to) possible worlds, as well as Max Cresswell (1990) for an argument against a Forbes-like position.

35 However, if one characterizes the notion of *de jure* rigidity in terms of an expression being rigid "in virtue of the semantical rules of the language," then, given that the semantical rules of the language state that "actual" is a rigidifying operator, actualized descriptions *will* count as *de jure* rigid designators (cf. Almog, 1986, pp. 223 ff.). If *de jure* rigidity is so characterized, then the *de jure/de facto* distinction is simply irrelevant to the question of whether names have descriptive content.

36 However, Kaplan does add qualifications (1989a, p. 497).

37 See also, in this context, Kripke's discussion of "Cicero" and "Jack the Ripper" in (1980, p. 79). For an interesting challenge to the whole idea of a descriptive name, see Bostock (1988).

38 To make issues of rigidity relevant for arguments against descriptive accounts of content, one needs to argue that proper names are, in the sense of Evans (1985b), *deeply rigid designators*, where an expression e counts as a deeply rigid designator of an object o just in case, for every possible world w, e refers to o when considered as *uttered in w*. Actualized descriptions are thus not deeply rigid designators. There are few attempts to address the question of whether names are deeply rigid; though see Deutsch (1989). Thanks to Sanford Shieh for discussion here.

39 See, for example, the appendix to ch. 5 of Dummett (1981a), appendix 3 of his (1981b), and ch. 2 of his (1991). The arguments outlined in the final two sections have their sources in these passages.

40 There is a third possibility, that is, that the occurrences of 'that' are unstructured names of the preceding sentence-tokens. But in this case it is even more difficult to see how "t" could take wide scope with respect to the modal operator in the next sentence. The only way I can see to defend Dummett's positive account is by using Kaplan's (1986) device of "arc quotes," maintaining that the "that" is replaceable by arc-quote names of the preceding sentences, which do license quantifying in.

41 See Dummett (1991, p. 48). By "formal semantic theory," I mean the project Robert Stalnaker calls "descriptive semantics" (see Chapter 35, REFERENCE AND NECESSITY).

42 I have characterized assertoric content as applying primarily to utterances rather than to sentence-types. But we can, from this characterization, obtain an equivalence relation of sameness of assertoric content which holds between sentence-types. Say that two sentence-types, S and S′, have the same assertoric content, just in case, for every normal context c, utterances of S and S′ in that context have the same assertoric content (for the notion of "normal context," see §I of my (1997)).

43 This is precisely Lewis's central conclusion (1981), albeit phrased in terms of Dummett's distinction between assertoric content and ingredient sense, rather than Lewis's vocabulary of "proposition" versus "semantic value" (1981, p. 95). Some of these facts have also been recognized (though used for different purposes) by Richard (1981; 1982) and Salmon (1986, ch. 2).

44 Ordinary language examples can, however, occasionally mislead here. For instance, the sentence "John believes something that was true yesterday, and false today" is perfectly acceptable. Yet the existence of such examples should not be taken as undermining the philosophical position that the objects of belief must be true or false absolutely. Such examples can be dealt with, as in Forbes (1989, p. 163), by interpreting the quantification substitutionally.

45 Similarly:

(a) The sun is shining. That's true now, but it won't be true tomorrow.
(b) The sun is shining. That's true here, but it's not true in Scotland.

46 According to Forbes (1989, part II), whereas assertoric contents are to be identified with Fregean thoughts, states of affairs are the objects of modal evaluation. My own view (Stanley, 1997) is that modal semantic value comes from the speech act of supposition, rather than assertion.

47 The classical 'Russellian proposition' view of content, for example, Kaplan (1989), Salmon (1986), and Soames (1987), is one view which entails RT.

48 Robert Stalnaker and Timothy Williamson deserve the greatest thanks for helping me with this chapter. In addition, extensive comments by Bob Hale and Crispin Wright substantively improved the chapter. I would also like to thank Michael Dummett, Kit Fine, Richard Heck, and Susanna Siegel for discussion.

References

Almog, J. 1986. "Naming without necessity." *Journal of Philosophy*, 83(4): 210–242.

Almog, J., J Perry, and H. Wettstein, eds. 1989. *Themes from Kaplan*. Oxford: Oxford University Press.

Barcan, R. 1947. "The identity of individuals in a strict functional calculus of second order." *Journal of Symbolic Logic*, 12(1): 12–15.

Barcan Marcus, R. 1961. "Modalities and intensional languages." *Synthese*, 13(4): 303–322.
Bostock, D. 1988. "Necessary truth and a priori truth." *Mind*, 97(387): 343–379.
Carnap, R. 1988 (1947). *Meaning and Necessity*. Chicago: University of Chicago Press.
Church, A. 1943. "Review of Quine's 'Notes on existence and necessity.'" *Journal of Symbolic Logic*, 8(1): 45–47.
Cresswell, M. J. 1990. *Entities and Indices*. Dordrecht, Netherlands: Kluwer.
Davies, M., and L. Humberstone. 1980. "Two notions of necessity." *Philosophical Studies*, 38(1): 1–30.
Deutsch, H. 1989. "On direct reference." In Almog, Perry, and Wettstein, 1989, pp. 167–195.
Dummett, M. 1981a. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.
Dummett, M. 1981b. *The Interpretation of Frege's Philosophy*. Cambridge, MA: Harvard University Press.
Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge: Harvard University Press.
Evans, G. 1982. *The Varieties of Reference*. Oxford: Clarendon Press.
Evans, G. 1985a. "Does tense logic rest upon a mistake?" In *Collected Papers*, pp. 343–363. Oxford: Clarendon Press.
Evans, G. 1985b. "Reference and contingency." In *Collected Papers*, pp. 178–213. Oxford: Clarendon Press.
Fine, K. 1989. "The problem of *de re* modality." In Almog, Perry, and Wettstein, 1989, pp. 197–272.
Fitting, M. 1993. "Basic modal logic." In *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 1, pp. 365–448. Oxford: Clarendon Press.
Fitting, M. 1991. "Modal logic should say more than it does." In *Computational Logic: Essays in Honor of Alan Robinson*, edited by J.-L. Lassez and G. Plotkin, pp. 113–135. Cambridge, MA: MIT Press.
Forbes, G. 1989. *Languages of Possibility*. Oxford: Blackwell.
Garson, J. 1984. "Quantification in modal logic." In *Handbook of Philosophical Logic*, edited by D. M. Gabbay and F. Guenthner, pp. 249–307. Dordrecht, Netherlands: Reidel.
Gibbard, A. 1975. "Contingent identity." *Journal of Philosophical Logic*, 4(2): 187–221.
Hodes, H. 1984a. "Axioms for actuality." *Journal of Philosophical Logic*, 13(1): 27–34.
Hodes, H. 1984b. "On modal logics which enrich first-order SS." *Journal of Philosophical Logic*, 13: 423–454.
Kaplan, D. 1986. "Opacity." In *The Philosophy of W. V. Quine*, edited by L. E. Hahn and P. A. Schilpp, pp. 229–289. La Salle, IL: Open Court.
Kaplan, D. 1989a. "Demonstratives." In Almog, Perry, and Wettstein, 1989, pp. 481–563.
Kaplan, D. 1989b. "Afterthoughts." In Almog, Perry, and Wettstein, 1989, pp. 567–614.
Kripke, S. 1963. "Semantical considerations on modal logic." *Acta Philosophica Fennica*, 16: 83–94.
Kripke, S. 1980 (1970). *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lewis, D. 1981. "Index, content, and context." In *Philosophy and Grammar*, edited by S. Kanger and S. Ohman, pp. 79–100. London: Reidel.
Montague, R. 1974. "Pragmatics and intensional logic." In *Formal Philosophy*, pp. 119–147. New Haven, CT: Yale University Press.
O'Brien, L. 1995. "Evans on self-identification." *Noûs*, 29(2): 232–247.
Quine, W. V. O. 1943. "Notes on existence and necessity." *Journal of Philosophy*, 40(5): 113–127.
Quine, W. V. O. 1947. "The problem of interpreting modal logic." *Journal of Symbolic Logic*, 12(2): 43–48.
Quine, W. V. O. 1956. "Quantifiers and propositional attitudes." *Journal of Philosophy*, 53(5): 177–187.
Quine, W. V. O. 1977. "Intensions revisited." In *Midwest Studies in Philosophy*, vol. 2, pp. 5–11. Morris, MN: University of Minnesota Press.
Richard, M. 1981. "Temporalism and externalism." *Philosophical Studies*, 39(1): 1–13.
Richard, M. 1982. "Tense, propositions, and meanings." *Philosophical Studies*, 41(3): 337–351.
Richard, M. 1987. "Quantification and Leibniz's law." *Philosophical Review*, 96(4): 555–578.
Salmon, N. 1982. *Reference and Essence*. Oxford: Blackwell.
Salmon, N. 1986. *Frege's Puzzle*. Cambridge, MA: MIT Press.
Segerberg, K. 1973. "Two-dimensional modal logic." *Journal of Philosophical Logic*, 2(1): 77–96.

Smullyan, A. 1947. "Review of Quine's 'The problem of interpreting modal logic.'" *Journal of Symbolic Logic*, 12(4): 139–141.

Smullyan, A. 1948. "Modality and description." *Journal of Symbolic Logic*, 13(1): 31–37.

Soames, S. 1987. "Direct reference, propositional attitudes, and semantic content." *Philosophical Topics*, 15(1): 47–87.

Soames, S. 1995. "Revisionism about reference: a reply to Smith." *Synthese*, 104(2): 191–216.

Stalnaker, R. 1977. "Complex predicates." *The Monist*, 60(3): 327–339.

Stalnaker, R. 1995. "The interaction of modality with quantification and identity." In *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus*, edited by W. Sinnott-Armstrong, in collaboration with D. Raffman and N. Asher, pp. 12–28. Cambridge: Cambridge University Press.

Stanley, J. 1997. "Rigidity and content." In *Logic, Language, and Reality. Essays in Honour of Michael Dummett*, edited by R. Heck, pp. 131–156. Oxford: Oxford University Press.

Strawson, P. 1959. *Individuals*. London: Methuen.

van Benthem, J. 1983. *Modal Logic and Classical Logic*. Napoli, Italy: Grafitalia.

37

Two-Dimensional Semantics

CHRISTIAN NIMTZ

1 2D Semantics: Ideas, Interpretations, and Issues

The theories that form the heterogeneous family of two-dimensional or 2D semantics are rooted in the tradition of possible-worlds semantics made popular by Saul Kripke and David Lewis. This semantics recognizes a *dependence of truth on fact*. Consider the sentence 'Pavarotti is famous.' Its truth-value depends on how things are. If things are as the sentence says, it is true; if they are not, it is false. A semantics may thus capture the truth-conditions of the sentence by specifying how its truth-value varies with the ways things could be. Equating ways things could be with possible worlds, we arrive at a core idea of intensional semantics – the idea that the truth-conditions of a sentence are (or weaker: can be modeled as[1]) its truth-values at all possible worlds. This idea generalizes. We can equate the semantic value of any expression with an intension, that is, a function from possible worlds to extensions, be it an individual (as in the case of singular terms), a set (as in the case of predicates), or a truth-value (as in the case of sentences).

Advocates of 2D semantics agree that recognizing a dependence of truth on fact is not enough. They urge us to acknowledge another dependence, which I, with deliberate looseness, call the *dependence of truth on meaning-fixing conditions*. Proponents of 2D accounts thus agree that the extensions of our expressions are doubly dependent on meaning-fixing conditions *and* on the respective facts – although they fervently disagree on what the former dependence amounts to.

Here is a way to motivate this idea. Suppose Pavarotti utters 'I am famous.' The truth-value of this indexical utterance at a possible world *w* depends on whether or not Pavarotti is famous at this world *w*. Yet its truth-value also depends on who made the utterance in the first place. Here we have a double dependence of the truth-value of an utterance at a world *w* on the facts in *w and* on the relevant content-fixing conditions – on whether or not Pavarotti is famous, and on who made the utterance in the first place.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

Advocates of two-dimensional semantics share a second trademark idea. They agree that the familiar apparatus of worlds-cum-intensions can be modified so as to capture both the dependencies they see. The resulting 2D apparatus discerns a first or 'A-dimension' of possible worlds playing the role of meaning-fixing conditions from a second or 'C-dimension' of possible worlds standing in as the relevant facts – hence *two*-dimensionalism.[2] It moreover assigns intensions in a way that captures the double dependence of extension on relevant facts and meaning-fixing situations.

Two-dimensional theories and ideas have influentially been proffered by David Kaplan (1977; 1989), Robert Stalnaker (1978; 2004), Gareth Evans (1979), Davies and Humberstone (1980), David Lewis (1980), Frank Jackson (1998; 2004; 2010), and David Chalmers (2002a; 2004; 2006). 2D semantics has been inspired by more formally oriented work in two-dimensional logic such as, for example, Åqvist (1973), Segerberg (1973), and van Fraassen (1977); see Kuhn (2012) and Humberstone (2004) for an overview. The diverse 2D accounts share little beyond the characteristic formal apparatus of worlds-cum-intensions that results from their trademark ideas. One reason for this is quite clear. Advocates of 2D semantics take different attitudes on what I dub 'orthodox Kripkeanism.'

Orthodox Kripkeans maintain that Kripke is right in his semantical doctrines concerning rigid designation and necessities *a posteriori*, and that he is right in his meta-semantics[3] leading to an anti-Fregean picture of reference and communication. Advocates of 2D semantics as a rule accept Kripke's semantic doctrines. (David Lewis is a notable exception to this.) They expect a semantics to make good on Kripkean claims such as, say, that 'Venus' designates rigidly, or that 'Water is H_2O' is necessary *a posteriori*. However, advocates of 2D theories take different attitudes towards Kripkean meta-semantics. There are prominent 2D theorists, such as David Kaplan and Robert Stalnaker, who embrace the orthodox Kripkean picture of reference and communication and employ 2D techniques in order to add to it. I will call anyone who likewise employs 2D techniques a '2D Kripkean.' Other 2D theorists, most notably Frank Jackson and especially David Chalmers (see also Davies and Humberstone, 1980), enlist 2D ideas to reject Kripkean meta-semantics. Jackson and Chalmers accept rigid designation and *a posteriori* necessities. But they reject Kripkean meta-semantics in favor of a Fregean, sense-based theory leading to a substantially different picture of reference and communication. I will call everyone who likewise employs 2D techniques a '2D Fregean.'[4]

Often somewhat misleadingly labeled 'two-dimensionalism,' the revisionist 2D Fregean program of Jackson and Chalmers has far-reaching implications for meta-philosophy and metaphysics. Jackson and Chalmers embrace a meta-philosophical picture that assigns conceptual analysis center stage (see Jackson, 1998). They can do so because their 2D Fregeanism allows for conceptual analysis, whereas Kripkeanism is inimical to it – if the semantic properties of, say, 'belief' are fixed causal-historically, there is little point in mulling over the concept in the armchair if you want to ascertain truths about beliefs.[5] This is presumably why the contemporary debate focuses on 2D Fregeanism, rather than on 2D semantics more broadly conceived (see García-Carpintero and Macià, 2006; Soames, 2005, and the contributions to the 2004 *Philosophical Studies* (118:1) volume on 2D semantics). In fact, much of the debate concerns two issues:

- Can 2D Fregeans devise a compelling semantics that is fundamentally Fregean, yet respects Kripke's semantic insights, and accommodates Kripke's arguments widely assumed to show that any sense-based account fails for proper names and natural kind terms?
- Is there any need for a 2D Fregean revision of orthodox Kripkeanism to begin with?

Within the philosophy of language, there is comparatively little controversy about the 2D accounts of Kaplan and Stalnaker. The former, especially, appears to be widely accepted as a standard semantics of indexicals.

2 The 2D Apparatus of Worlds-cum-Intensions

The commitments of all 2D theories can basically be captured by one and the same formal framework of worlds-cum-intensions. I call it 'the 2D apparatus.' To understand the 2D apparatus, then, is to understand the general structure that all 2D accounts share.

In order to capture the dependence of truth on fact, traditional intensional semantics assigns every expression an intension, that is, a function *f*: $W \rightarrow E$ from worlds to extensions, as its semantic value. Consider the sentence 'Pavarotti is famous.' Its intension is a function *f*: $W \rightarrow$ {*true*, *false*} from worlds to truth-values such that the sentence gets assigned the truth-value *true* at a world *w* just in case Pavarotti is famous in *w*. This intension is determined by the intensions of the name and the predicate the sentence comprises. The intension of the name 'Pavarotti' is a function *f*: $W \rightarrow$ D from worlds *w* to individuals (intuitively: the individual that is Pavarotti in *w*), and the intension for 'is famous' is a function *f*: $W \rightarrow P$(D) from worlds *w* to sets of individuals (intuitively: the set of individuals famous in *w*).[6] Possible worlds here figure as counterfactual alternatives to our actual world, and they play the role of being what sentences are evaluated at for truth or falsity. Let us label worlds playing this role 'C-worlds.'

Proponents of two-dimensional semantics maintain that C-worlds are not enough. In order to capture the dependence of truth on meaning-fixing conditions, we need a prior dimension of worlds I call 'A-worlds.' To again use indexicals for illustration, consider Pavarotti uttering 'I am famous.' Call this utterance *u*. Whether *u* is true at some C-world *w* depends on the facts at *w*. Yet it also depends on who uttered *u* in the first place. Given that it was Pavarotti who produced *u*, *u* is true at some C-world *w* just in case Pavarotti is famous in *w*. If someone else uttered *u* instead, the truth-conditions of the utterance will vary accordingly. We thus find that the intension expressed by an utterance of 'I am famous' systematically depends on prior meaning-fixing conditions, or A-worlds: it depends on the context of utterance. There is a two-stage procedure at work. First of all, the context of utterance (= the A-world) fixes *u*'s truth-conditions. It depends on the respective A-world which specific function *f*: $W_C \rightarrow$ {*true*, *false*} from C-worlds to truth-values is *u*'s intension. That Pavarotti figured as the speaker in the context of utterance of *u* led to *u* having its specific intension. Second, given that *u* has this intension, it can be evaluated for truth or falsity by considering the relevant facts (= a C-world): *u* is true at some C-world *w* just in case Pavarotti is famous in *w*.

All two-dimensional semantics envisage such a two-step procedure of evaluation. They therefore distinguish A-worlds standing in as meaning-fixing conditions from C-worlds serving as the relevant facts. It is most natural (though controversial) to hold that the same possible worlds figure as A- and C-worlds, respectively; they just play different roles in these different capacities. There is one important exception to this. In order for the account to do the work it is supposed to do, A-worlds need to be *centered*, that is, they must come with a *center* consisting of a speaker (if any), an audience (if any), a place, a time, and so on, highlighted. (I throughout assume A-worlds to be centered. If need be, I use 'w^*' to emphasize that we are dealing with the world *w* with some arbitrary center marked.)

Second dimension
C-worlds →

First dimension
A-worlds (centered) →

	w1	w2	w3
w1*	t	t	f
w2*	f	t	t
w3*	t	t	f

Figure 37.1 The matrix specifies all three intensions of a sentence for a small sample of worlds. Every row specifies a C-intension. The diagonal specifies the A-intension. The matrix as whole specifies the 2D intension.

The resulting apparatus with its two dimensions of worlds allows us to distinguish three kinds of intensions. First of all, there are *C-intensions*. (Kaplan calls these 'contents'; Chalmers speaks of 'secondary' or '2-intensions.') A C-intension is a function f: $W_C \rightarrow E$ from C-worlds to extensions. A C-intension of a sentence δ specifies for any C-world w whether δ is true or false at w. Stalnaker dubs C-intensions of sentences 'horizontal propositions' since they corresponds to a row in the sentence's 2D matrix (see Figure 37.1). C-intensions are the intensions figuring in classic possible-worlds semantics.

Second, there are *A-intensions*. (Chalmers calls these 'primary intensions' or '1-intensions.') An A-intension is a function f: $W_A \rightarrow E$ from A-worlds to extensions. The A-intension of a sentence δ specifies for any A-world w whether, given that w figures as the meaning-fixing conditions for δ, δ is true or false when evaluated at the very same world w. Since the A-intension of a sentence corresponds to the diagonal in the sentence's 2D matrix (see Figure 37.1), Stalnaker dubs propositional A-intensions 'diagonal propositions.'[7]

Finally, there are two-dimensional intensions. (Kaplan speaks of 'characters.') The 2D intension of an expression α specifies for any A-world w which C-intension α has, given that w figures as the meaning-fixing conditions. 2D intensions thus are functions f: $W_A \rightarrow$ (f: $W_C \rightarrow E$) from A-worlds to C-intensions; alternatively, we can take them to be functions f: $W_A \times W_C \rightarrow E$ from pairs of A- and C-worlds to extensions. The whole schema given in Figure 37.1 specifies (part of) a 2D intension.

As emphasized above, advocates of 2D semantics agree that the apparatus of worlds-cum-intensions can be modified so as to capture both the dependencies they see. As should now be apparent, the modification consists in distinguishing A-worlds from C-worlds. It should also be apparent how the resultant 2D apparatus captures the dependence of truth on fact and the dependence of truth on meaning-fixing conditions (as well as the double dependence of truth on both). It does so by way of C-intensions, A-intensions, and 2D intensions, respectively.

The various 2D accounts tend to consider one of these intensions as fundamental, and the others as derivative. But on the face of it, it is a virtue of the general 2D schema that it encompasses all three kinds of intensions.[8]

First of all, discerning A- and C-intensions allows the 2D apparatus to account for conflicting pre-theoretic semantic judgments. Suppose Emma and Joe both utter 'I am fine.' Have they said the same thing? Intuitively speaking, yes and no. Both utterances are true

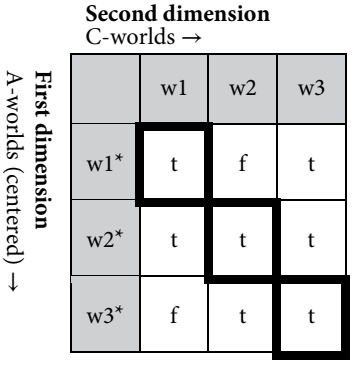

Figure 37.2 The matrix specifies (part of) the intensions of 'I am here.' It combines a necessary A-intension or diagonal with contingent C-intensions.

just in case the respective speaker is fine. They say the same thing by virtue of expressing the same A-intension. However, what Emma said is true just in case Emma is fine, whereas what Joe said is true just in case Joe is fine. So they do not say the same thing, by virtue of expressing different C-intensions.

Second, discerning A- and C-intensions allows us to account for an apparent modal puzzle. Suppose I utter 'I am here now.' To do so is to make a contingent claim. I could right now have been someplace else. However, I could not possibly have said something false with my utterance, and what could not possibly have been false is necessarily true. The 2D apparatus dissolves the tension. The utterance combines a contingent C-intension with a necessary A-intension. In Gareth Evans's (1979, p. 179) terminology, the statement is superficially contingent, yet deeply necessary. Generalizing, we may acknowledge two different modal dimensions and devise two different types of modal operators. For instance, we could rule that '$\Box\delta$' is true just in case δ's C-intension yields *true* for any C-world, whereas '$\boxdot\delta$' is true just in case δ's A-intension yields *true* for any A-world (see Kuhn, 2012, for details).

Finally, 2D intensions allow for a homogeneous account of semantic competence. The following three apparently innocuous ideas spell trouble once they are combined:

(i) Semantic competence concerns sentence-types.
(ii) To understand a sentence is to know its truth-conditions.
(iii) Sentence-types containing indexicals do not have truth-conditions, since indexical-types do not refer – there is no object the *type* 'I' refers to.

Here is a two-dimensional way to solve this puzzle. Let us trade (ii) for the idea that to understand a sentence is to know its 2D intension. This idea is natural enough for indexical sentence-types such as, for example, 'I am famous.' If the type 'I' does not refer, the sentence-type does not have truth-conditions. So my semantic competence cannot consist in knowing these. There still is something I know. I know which truth-conditions an utterance u of this sentence has, given that u occurs in a specific context (= A-world). My semantic competence with 'I am famous' can thus be captured by a 2D intension detailing a function f: $W_A \rightarrow (f\text{: } W_C \rightarrow E)$ from A-worlds to C-intensions. The same holds arguably and on some 2D accounts even trivially true of my competence with non-indexical sentence-types.

All varieties of 2D semantics employ basically the same 2D apparatus. There also is fundamental agreement on to how to understand the C-dimension. C-worlds are Kripke's

metaphysically possible worlds understood to be counterfactual alternatives to our actual world, and they are employed to capture the dependence of truth on fact. C-intensions are the intensions of traditional intensional semantics, and C-necessity is metaphysical necessity. There is no such consensus on the A-dimension; far from it. 2D Kripkeans like Kaplan and Stalnaker and 2D Fregeans like Jackson and Chalmers hold very different views on what the dependence of truth on meaning-fixing conditions amounts to, and why our semantics should capture it. Given the 2D apparatus, this is to say that they differ on what A-worlds are and why we need them, what A-intensions capture, and what A-necessity amounts to. A key reason for this is that these theorists take different attitudes on orthodox Kripkeanism.

3 Essential Background: Orthodox Kripkeanism

Orthodox Kripkeans, as I employ the term, embrace the key semantic and meta-semantic doctrines developed by Kripke (1971; 1980) and in work congenial to his (see Putnam, 1975; Burge, 1986; Kaplan, 1977). Orthodox Kripkeanism arguably is the presumed default position in much of contemporary semantics. It combines five doctrines of interest to us.

According to the first Kripkean doctrine, many of our terms designate rigidly. Glossing over intricacies (see Hughes, 2004, pp. 19–24), a term is a "rigid designator if in every possible world it designates the same object" (Kripke, 1980, p. 48). Kripkeanism considers proper names like 'Cicero' and natural kind terms such as 'gold' to be clear instances of rigidly designating expressions. Kaplan's work (see next section) has added indexicals such as 'today' to this list.

According to the second Kripkean doctrine, there are necessities *a posteriori*. Let us understand this idea modestly as stating that there are necessarily true yet *a posteriori* sentences. Read thus, the doctrine is agreeable to any self-proclaimed Kripkean. This does not hold true of the stronger reading according to which there are *propositions* that are necessary as well as *a posteriori* (see Soames, 2011, p. 79).[9] Stock examples of necessities *a posteriori* are true identity statements with rigid designators flanking the identity sign such as, for example, 'Cicero = Tully,' or 'Water = H_2O.'

Although it arguably figures in Kripke, the third doctrine is more closely associated with David Kaplan. According to this doctrine, many singular terms are 'directly referential' (Kaplan, 1977, p. 493) – they contribute their referent, rather than some descriptive content, to the contents expressed by sentences comprising them. Standard examples of directly referential expressions are the rigid designators listed above: proper names, natural kind terms, and indexicals. Suppose that 'Venus' is directly referential. Then all it contributes to the content expressed by 'Venus is a planet' is the celestial body it refers to. Any directly referential expression designates rigidly. The converse does not hold. The description 'the actual inventor of the zip' picks out one and the same individual in all possible worlds, *viz.* whoever invented the zip in our actual world. But it arguably contributes the descriptive property *having actually invented the zip* to contents expressed by the statements it occurs in. Orthodox Kripkeans are thus prone to contrast proper names and natural kind terms with expressions comprising rigidifying devices such as 'actually,' or Kaplan's (1989, pp. 579–582) 'dthat.'[10] All these terms designate rigidly. But only the former qualify as directly referential.

The fourth doctrine translates the idea of direct reference into a view of propositions. According to this doctrine, singular statements comprising directly referential expressions express singular propositions. A singular proposition is composed of objects and properties,

and it can be symbolized by an *n*-tuple of these. Given that 'Venus' is directly referential, 'Venus is a planet' expresses the singular proposition < Venus, *being a planet* >. And given that 'here' refers directly, an utterance of 'Here it is nice' expresses < *o*, *being nice* > with *o* being the referent of 'here' in the context of utterance.[11]

Orthodox Kripkeanism adds a characteristic meta-semantics to these four semantic doctrines, yielding a distinctive picture of reference and communication. According to this meta-semantics, reference-fixing for proper names and natural kind terms is extra-semantic. What 'Cicero' or 'water,' respectively, refer to is not even partly determined by a sense-like semantic property these terms have. Their reference is entirely fixed by mechanisms feeding on non-semantic properties of the terms. Suppose our linguistic ancestors introduced 'Cicero' as a designator for the Roman orator who denounced Catiline, and 'water' as a designator for the watery substance of our acquaintance (to employ a convenient abbreviation). Orthodox Kripkeans insist that the descriptive properties *being the Roman orator who denounced Catiline* and *being the watery substance of our acquaintance* function as mere ephemeral auxiliaries; the same holds true for any other associated descriptive property. *How* anyone picked out the referents is incidental. All that is of semantic importance is *which* objects are assigned as referents.

A famous instance of this general idea is the 'causal-historic' account of reference as proposed by Kripke (1980) and Putnam (1975). On this view, the referent of a proper name such as 'Cicero,' as used in a community, is the very object assigned as referent at the historical origin of this use of the term (see Kripke, 1980, p. 96 f.; Putnam, 1975). Analogously, a natural kind predicate such as 'water,' as used in a community, applies to the items or samples that are of the same natural kind as the items or samples employed as paradigms at the historical origin of this use of the term. On the causal-historic account, then, the referent of a name or kind term α thus is "determined by a 'causal' chain of communication rather than a description" (Kripke, 1980, p. 59, fn. 22). It is thus fixed quite independently of any semantic property α already has.

Kripkeanism marks a radical break with the Fregean tradition in semantics. Fregeans insist that any expression α has as its fundamental semantic property a descriptive content or 'sense' (*Sinn*) capturing α's cognitive significance by spelling out how the term's referent is presented, identified, or thought of. Paradigmatically, the sense of a name such as 'Cicero' is a descriptive content that can be expressed by a definite description such as, let us assume, 'the Roman orator who denounced Catiline.' The sense of an expression plays a complex role. Sense (i) determines reference, (ii) captures cognitive significance, (iii) is grasped by competent speakers, and (iv) is part of what is believed, said, and communicated. The referent of 'Cicero' (if there is one) is the unique individual satisfying the associated condition of *being the Roman orator who denounced Catiline*, and it is by this description that speakers employing the term identify or think of its referent. To understand a term thus is to knowingly associate the (or at least: a) conventionally assigned sense with it. By the same token, to understand a sentence amounts to grasping the descriptive content, called a 'thought' (*Gedanke*) by Frege, composed of the senses of the expressions that make up the sentence. It is thoughts that are expressed and communicated by the (assertoric) utterances we perform.

Orthodox Kripkeans agree that this picture may, *modulo* socio-linguistic phenomena like Putnam's (1975, p. 227) 'division of linguistic labor,' be correct for descriptive terms such as 'grandmother.' But they consider it fundamentally flawed as a general account of meaning. Kripke (1980) influentially argues that any sense-based account of reference-fixing and understanding (and, by consequence, communication) fails for proper names and natural

kind terms. Orthodox Kripkeans conclude that these terms *simply do not have senses*. Their reference is fixed extra-semantically rather than by satisfaction, they designate rigidly and refer directly, and sentences such as 'Water is transparent' express singular propositions rather than Fregean thoughts. By consequence, competence with such terms cannot consist in grasping their senses, and what we semantically assert and communicate with (assertoric) utterances of sentences comprising them cannot be Fregean thoughts. These terms do not have senses, and these sentences do not express Fregean thoughts.

Advocates of 2D semantics as a rule accept Kripke's ideas concerning rigid designation and necessities *a posteriori*. This semantic consensus comes with a deep rift on meta-semantics. 2D Kripkeans like Kaplan and Stalnaker embrace the Kripkean meta-semantics and the picture of reference and communication that comes with it. 2D Fregeans like Jackson and Chalmers reject the Kripkean meta-semantics in favor of an account that explains semantic properties in terms of underlying descriptive meanings playing the roles of Fregean senses. In the context of 2D theories, this rift translates into different views on what the dependence of truth on meaning-fixing conditions comes to, and what the A-dimension is good for.

4 Kaplan's 2D Semantics for Indexicals

David Kaplan's (1977; 1989) semantics aims to cover pure indexicals such as 'today' (on which I will focus) alongside impure indexicals such as 'you' and demonstratives such as 'this.' Kaplan starts off with two 'obvious principles' (Kaplan, 1977, p. 492):

(i) The referent of an indexical systematically depends on the respective context.
(ii) Indexicals are directly referential.

The first principle looks innocuous enough. Who doubts that the referent of, say, 'today' varies with the day in question? Kaplan takes the second principle to be just as incontrovertible. He stresses that we evaluate utterances such as Pavarotti's 'I am famous' as saying *of* some individual that he is famous. All an indexical as used in a context does is to pick out some individual (person, date, time, place, etc.), and it is the individual picked out, rather than some way to identify it, that affects what is said and thus enters into the proposition expressed. However, Kaplan embraces a third idea that does not sit too easily with his principles:

(iii) 'Indexicals, in general, have a rather easily statable descriptive meaning.' (Kaplan, 1977, p. 498)

For example, we understand 'today' to mean something like 'the present day.' Kaplan even agrees that these descriptive meanings mediate reference and are known to competent speakers. They thus perform key functions of Fregean senses. But how can indexicals have sense-like meanings *and* be directly referential? Kaplan's 2D semantics for indexicals convincingly squares these two features of indexical expressions.

Kaplan's semantics pivots on two distinctions. First of all, he distinguishes "possible occasions of use – which I call *contexts* – from possible circumstances of *evaluation* of what was said on a given occasion of use" (Kaplan, 1977, p. 494). This is a distinction between

two kinds of possible situations. The former figure as those situations in which an utterance can be made. The latter figure as those situations at which what is said by a sentence in a context can be evaluated. Second, Kaplan distinguishes *contents* from *characters*.[12] This is a distinction between semantic properties. Contents are semantic properties of occurrences, whereas characters are semantic properties of types. The content of a sentence in a context *c* is what is said with that sentence in *c*. Likewise, the content of a sub-sentential expression in a context *c* is what it contributes to what is said in *c*. For indexical sentences, the content expressed will depend on features of the context the sentence occurs in. A character is a conventionally assigned linguistic rule that descriptively specifies, for any context *c*, which content an occurrence of the respective type expresses in *c*. We may think of the character of 'I' as the rule "I,' as used in any context *c*, refers to the speaker or writer in *c*' (see Kaplan, 1977, p. 505). And we may think of the character of 'today' as the rule "today,' as used in a context *c*, refers to the day of the context *c*.'

Kaplan's account neatly fits the 2D apparatus sketched above. Think of circumstances of evaluation as C-worlds, and of contexts of use as A-worlds. Following Kaplan, we may model contents as functions from circumstances to extensions. This makes them C-intensions. And we may conceive of characters as functions from contexts to contents. This makes them 2D intensions. The Kaplanian picture we arrive at, then, is this. Any sentence-type δ has a compositionally determined character. The character of δ is a 2D intension that specifies what is said with an occurrence of δ, that is, which content δ expresses for any context *c* (= A-world). The contents thus expressed are C-intensions, that is, function from circumstances of evaluations (= C-worlds) to truth-values. When δ comprises a directly referential term, this term receives a constant object of evaluation at any circumstance of evaluation. Which object this is is already fully settled by the context, and thus "simply independent of the circumstance" (Kaplan, 1977, p. 497). This is why Kaplan equates the contents expressed by utterances of such sentences more narrowly with singular propositions.

For illustration, consider the sentence 'I am famous.' Its character (=2D intension) specifies for any context of use *c* what is said with an occurrence of 'I am famous' in *c*. Given that 'I' is directly referential, as Kaplan maintains, any occurrence of 'I am famous' says *of* the speaker or writer in *c* that (s)he is famous. It thus expresses a content (= C-intension) that is true at some circumstance of evaluation (= C-world) just in case this person is famous in that C-world. Consider a context c_1 where Pavarotti utters 'I am famous.' The character of the sentence-type together with the fact that Pavarotti is the speaker in c_1 determines the content expressed, and this content is true at a circumstance of evaluation just in case Pavarotti is famous at that C-world. Since the context fully settles that Pavarotti is the unvarying object of evaluation, Kaplan thinks of the content expressed more narrowly as the singular proposition <Pavarotti, *being famous*> containing Pavarotti himself as a constituent.

Kaplan's account combines explanatory power with theoretical clarity. First, Kaplan's semantics neatly explains how indexicals can have sense-like descriptive meanings *and* be directly referential: indexical *types* have sense-like meanings in the form of characters, whilst their *occurrences* in contexts are directly referential. The character of an indexical α fixes what is said with an occurrence of α in a context. But it does not enter into the content thus expressed. Second, Kaplan's theory elegantly accounts for modal puzzles such as the contingency of the apparently necessary 'I am now here.' An utterance of 'I am here' by me now expresses a singular proposition comprising me, *this* time, and *this* place. This proposition

clearly is contingent – there are C-worlds where it is false. But the character of 'I am here' apparently guarantees that any utterance of it is true in the context in which it is uttered. Third, Kaplan's semantics explains how indexical utterances function in communication. I know that Pavarotti made a *de re* claim about Pavarotti by uttering 'I am famous,' since I know the character of 'I' and since I know of Pavarotti that he just spoke. I know the former by being a competent speaker, and the latter by being a perceptive observer. Fourth, Kaplan has a clear answer to what the dependence of truth on meaning-fixing conditions comes to, and what A-worlds are good for. The former comes down to ordinary context-dependence, and we need A-worlds to model contexts of use. (Kaplan has little use for A-intensions; his picture absorbs these into characters.)

Subscribing to the orthodox Kripkean account of proper names and natural kind terms, Kaplan intends his 2D theory to add to the orthodox Kripkean case. But Kaplan's work arguably transforms the Kripkean picture. To begin with, since characters are composite and since any non-indexical expression can be combined in sentences with an indexical, all expressions get assigned characters. The difference between 'I' and 'Pavarotti' is not that the former has, whereas the latter lacks a character. The difference is rather that the content expressed by 'I' varies with the context, whereas that of 'Pavarotti' does not. What is more, Kaplan adds descriptivist elements to a Kripkean theory. On Kaplan's semantics, indexicals designate rigidly, yield necessities *a posteriori* (consider 'Today = September 21, 2015,' as uttered on September 21, 2015), and refer directly. But their reference in a context is fixed by a character – and characters are sense-like descriptive meanings conventionally associated with these expressions.

This yields a worry. Why should we think of Kaplan's semantics as a Kripkean account, rather than as a sophisticated Fregean theory? To defuse this challenge, Kaplan insists that in his theory, content rather than character is the rightful heir to Fregean sense (see Kaplan, 1989, p. 568). He explains: "[T]he issue is not whether the information used to determine the referent is descriptive or not. It is rather whether the relevant information, in whatever form, *is a part of what is said*" (1989, p. 578; my italics). What Pavarotti expresses with an utterance of 'I am famous' is just the singular proposition < Pavarotti, *being famous*>. No thought-like descriptive meaning gets expressed. Kaplan here subscribes to an idea I call the *preeminence of C-intensions*: Contents (= C-intensions) are the exclusive semantic vehicles of linguistic communication. What gets semantically expressed, asserted, or communicated are C-intensions. Descriptive meanings play a role in communication only in so far as they happen to be C-intensions, as is the case with purely descriptive sentences. The preeminence of C-intensions is a deeply orthodox Kripkean idea. It is inspired by the Kripkean conviction that proper names, natural kind terms, and other directly referential expressions do not have descriptive meanings to contribute to what is said.

The preeminence of C-intensions arguably is an independently motivated commitment Kaplan adds to, rather than extracts from, his semantics. This feels problematic. On Kaplan's 2D semantics, every term has a character. But A-intensions are easily recoverable from characters. By consequence, in so far as competent speakers know the characters of their expressions, they should know the A-intensions their sentences semantically express. So, going by Kaplan's own semantics, why shouldn't we expect utterances of sentences to standardly express and communicate A-intensions over and above C-intensions? Even more worrisome to an orthodox Kripkean, we may well take Kaplan's treatment of indexicals as a blueprint to explain how names and natural kind terms can have full-blown Fregean senses even though they refer directly. This would allow us to acknowledge Kripke's semantic

doctrines, yet reject the Kripkean picture of reference-fixing. Such a 'generalized Kaplan paradigm' (Stalnaker, 2004, p. 309) would be a Fregean rival to orthodox Kripkeanism.

5 Robert Stalnaker, or 2D Pragmatics

Robert Stalnaker also commits to the orthodox Kripkeanism picture according to which C-intensions are the exclusive semantic vehicles of linguistic communication. He prominently acknowledges, however, that this picture faces challenging puzzles (see Stalnaker, 1978; 2004; 2014). Most pressingly, its champions need to explain how sentences expressing non-informative C-intensions can be informatively asserted, as they manifestly are. Stalnaker relies on the 2D apparatus to propose a pragmatic solution to these puzzles (see Stalnaker, 1999, pp. 12–19, for an overview).

Here is a claim about information: To be informative (to carry information, to represent) is to 'exclude possibilities' (Stalnaker, 2004, p. 300; see Jackson, 2001, p. 617). A statement is informative in a context *c* only if the proposition it expresses in *c* excludes some of the possibilities compatible with *c*. Should I choose to accept it, such a statement allows me to narrow the range of possible worlds that I, by the standards of the context, consider as candidates for the situation I actually am in. By the same token, to assert that *p* is to propose "to exclude from the possible situations compatible with the context those in which the proposition asserted is false" (Stalnaker, 2004, p. 300). Suppose you assert 'Today is Monday.' You here propose to exclude all the non-Monday-worlds from our common representation of how our present situation might be.

Although quite compelling, our little theory entails that informativity requires contingency. A statement *p* is informative only if there is at least one non-*p*-world. This leads to a puzzle. We clearly make informative assertions using statements such as 'Hesperus = Phosphorus,' or 'Water = H_2O.' But on orthodox Kripkean premises, statements such as these exclusively express necessary (singular) propositions that do not exclude any possibility. For example, 'Hesperus = Phosphorus' exclusively expresses < Venus, *being identical with*, Venus>. So we need to explain "how it is that a necessarily true statement could be used to convey contingent information?" (Stalnaker, 2004, p. 303).

Stalnaker's pragmatic resolution of this puzzle is premised on a causal theory of reference. To this, he adds two ideas. First, the meanings of proper names, natural kind terms, and the like are contingent in a specific manner. Consider a possible world *w* where the watery substance of our acquaintance is XYZ, rather than H_2O. Suppose we had introduced 'water' just as we actually did relying on the property *being the watery substance of our acquaintance*, but that *w* happened to be the actual world. Then our term 'water' had rigidly designated XYZ, rather than H_2O. The meanings of all the other terms governed by a causal-historic semantics do likewise depend on the respective actual world. We can employ the 2D apparatus to model this dependency. Let us understand A-worlds to be alternative *actual* worlds, that is, possible candidates for the world we actually inhabit. (If you like, A-worlds are counter*actual*, whereas C-worlds are counter*factual*.) Then the A-intension f: $W_A \rightarrow E$ of a term α specifies, for any A-world *w*, which extension α would have, given that *w* was actual. Thus understood, any A-intension is metalinguistic. It specifies what α would mean, rather than what it actually does mean.

Second, when an utterance violates a pragmatic maxim, we reinterpret. Suppose some utterance of 'It is raining' violates the Gricean maxim of relevance. We then look for the

alternative relevant content the speaker intends to convey with his utterance. What she might want to convey is that no, she does not want to go to the zoo. Stalnaker maintains that we do precisely the same when confronted with an assertion of a sentence expressing a non-informative necessary proposition such as 'Hesperus is Phosphorus.' (This of course presupposes that competent speakers recognize that 'Hesperus is Phosphorus' expresses a necessary and thereby non-informative proposition in the first place, and so see a need to re-intepret.) Here we look for an alternative informative proposition the speaker intends to convey. We do not have to look far. As explained above, the likes of 'Hesperus' have their C-intensions contingently. By consequence, the A-intension of a sentence such as 'Hesperus = Phosphorus' or 'Water is H_2O' will as a rule be contingent, too. So what we do by default is to *diagonalize*: we understand the speaker as intending to convey the contingent and therefore informative propositional A-intension or 'diagonal proposition' (Stalnaker, 1978, p. 318) the sentence has in the present context, rather than the necessary C-intension the sentence semantically expresses given the present context.

Stalnaker takes this "diagonalization strategy" (Stalnaker, 2008, p. 42) to provide a general solution to our puzzle. He also understands diagonalization to be a purely pragmatic maneuver. The A-intensions we enlist in diagonalization are not additional semantic values. They rather capture metalinguistic facts about the variability of meaning on a par with platitudes such as, say, 'Had we named Venus 'Mars,' 'Mars' would designate Venus.' So, although Stalnaker proffers a 2D account, he does not propose a 2D *semantics*. Stalnaker rather vigorously sticks to orthodox Kripkeanism, and he employs 2D techniques precisely in order to avoid committing to a two-dimensional semantics.

The success of this maneuver can be questioned, however. Stalnaker introduces diagonalization as a general strategy. But for any normal context *c* in which some directly referential term may plausibly occur, this term may plausibly figure in *c* in a sentence expressing a necessary C-intension. Suppose you are talking about water. Then uttering 'By the way, water is H_2O, as I have recently learned' is fairly unremarkable. So diagonalization must basically cover all directly referential terms and all the normal contexts these can occur in. But to diagonalize within some context *c* requires that speaker and hearer share in *c* descriptive knowledge rich enough to uniquely identify the actual referent of the directly referential expression concerned. When confronted with the uninformative 'Water is H_2O,' for example, you need to enlist some descriptive property that the actual referent of 'water' uniquely satisfies, at least in the context. So if Stalnaker is right about diagonalization, he seems compelled to grant that competent speakers knowingly associate descriptive contents that uniquely identify referents with their directly referential terms (basically) across the board. Any Fregean will take this to be a major concession.

Stalnaker does not think so. He emphasizes that the descriptive knowledge associated with a directly referential term varies with the context, and he insists that it is pragmatic rather than semantic. But the actual extent of the variation is debatable, and any Fregean theory of reference, especially one built around a cluster-theory such as, for example, Searle's (1958) account of names, has us expect a good deal of contextual variation in what is saliently communicated anyhow. Yet if that is so, what justification is there to class the associated descriptive content as pragmatic rather than semantic?

In fact, Stalnaker's concession creates a more serious problem. It threatens to undercut the very Kripkean model of communication diagonalization was designed to salvage. Stalnaker assumes that standard communication exclusively involves C-intensions. Diagonal propositions (= A-intensions) are communicated only when normal communication

breaks down, or so Stalnaker maintains. But as we have seen, competent speakers are bound to knowingly associate A-intensions with their directly referential terms across the board, even though these might vary with the context. They also are bound to knowingly associate A-intensions with their descriptive terms across the board, since here A- and C-intensions coincide. But if speakers are aware of the contextually indicated A-intensions of their terms anyway, shouldn't we expect that utterances of sentences comprising directly referential terms standardly communicate contextually indicated A-intensions as well?

6 Enter the 2D Fregeans: Jackson, Chalmers, and the Primacy of A-Intensions

Jackson and Chalmers agree that the Kripkean meta-semantics is misguided and that those who assign C-intensions preeminence in reference and communication have things backwards. However, they accept Kripke's doctrines of rigid designation and necessities *a posteriori*. (They could acknowledge direct reference and singular propositions, too, but may well see no need to do so.) So Jackson and Chalmers devise a 2D Fregean semantics that accounts for these semantic doctrines. Jackson and Chalmers welcome the support their semantics lends to conceptual analysis. Their 2D Fregeanism explains what conceptual analysis aims at (it aims at ascertaining A-intensions), and it explains why conceptual analysis is feasible even for natural kind terms such as, for example, 'causation' (it is so feasible because even natural kind terms have A-intensions) (see Jackson, 1998, ch. 2). However, Jackson and Chalmers harbor different ideas about the status and aim of the 2D Fregean projects they pursue. They consequently take rather different views on why we should hold that A-intensions are fundamental, and in what sense knowledge of A-intensions is *a priori*.

The semantics Jackson and Chalmers devise is Fregean in that it assigns all our expressions or their occurrences a sense-like descriptive content that (i′) plays a key role in determining reference, (ii′) captures a kind of cognitive significance, (iii′) is grasped by competent speakers, and (iv′) is often part of what is said and communicated. The semantics Jackson and Chalmers devise is two-dimensional in that it assigns any expression or occurrence an A-intension and a C-intension (with the latter potentially varying with the actual world as described by a 2D intension). These are not on a par. Rather, A-intensions are fundamental in that they fit the profile laid out by (i′)–(iv′) and thus basically play the role of Fregean senses. More specifically, a 2D Fregean semantics assigns any expression an A-intension, that is, a function f: $W_A \rightarrow E$ from actual worlds to extensions, and a C-intension, that is, a function f: $W_C \rightarrow E$ from counterfactual worlds to extensions. The C-dimension is interpreted traditionally. C-worlds are understood to be worlds considered as counterfactual and equated with comprehensive metaphysical possibilities, and C-intensions are taken to be the familiar traditional intensions. The A-dimension, however, is now seen as genuine semantic, as well as fundamental. A-worlds are understood to be worlds 'considered as actual,' that is, alternative actual worlds, and A-intensions are understood to be descriptive contents. A-intensions here do not metalinguistically trace changes *in* meaning, as they do in Stalnaker's account. They rather capture the descriptive contents our terms have as we actually use them.

On the 2D Fregean picture, then, any term or occurrence has a descriptive content as given by its A-intension. These contents are fundamental in that they determine C-intensions. Here we need to distinguish two cases. Either the A-intension of a term α all by itself

determines α's C-intension. This holds true for terms like 'grandmother' whose C-intension does not vary with the actual world. The A- and C-intensions of these terms coincide (given that we ignore centering). Or the A-intension of a term α fixes α's C-intension together with the actual world. This holds true for terms like 'water' or 'the actual inventor of the zip' whose C-intensions do depend on the actual world, or, more specifically: on what the actual watery substance of our acquaintance happens to be, and on who actually invented the zip, respectively.

Applying this account to descriptive terms such as, for example, 'triangle' or 'grandmother' is straightforward. A descriptive property determining what 'grandmother' applies to is easy to find: 'grandmother' applies to female parents of parents. The term's A-intension thus is a function from A-worlds to sets of female parents of parents. This A-intension by itself determines the term's C-intension. Since 'grandmother' applies to female parents of parents regardless of how the actual world happens to be, both its A- and its C-intension mark the very same function from worlds to female parents of parents (again, given that we ignore centering).

In applying their ideas to names and natural kind terms, the 2D Fregeans in effect take their cue from Kaplan. Kaplan equates the character of an indexical α with a rule descriptively identifying the referent of α for any context, and α's content with whatever satisfies that rule in the respective context *c*. The former remains stable across contexts, whereas the latter may vary. Substituting actual worlds for contexts of use and taking the distinction between A- and C-intensions to concern types and tokens alike, the 2D Fregean applies her ideas to proper names or natural kind terms in much the same fashion. She equates the A-intension of some such term α with the role carved out by the descriptive property associated with α, and she equates the C-intension of α, given that some A-world *w* is actual, with whatever fills that role in *w*. The former remains stable across A-worlds, whereas the latter may vary. Suppose we equate the A-intension of the natural kind term 'water' with the role of *being the watery substance of our acquaintance*. Then this is what 'water' stably refers to across A-worlds. Given that some A-world *w* is actual, the term's C-intension then is whatever happens to fill that role in *w*. 2D Fregeans thus hold that names and natural kind terms are *actuality-dependent* by virtue of their meaning – it is built into the semantics of 'water' that the term rigidly designates H_2O *if* the actual world is one where H_2O is the watery substance of our acquaintance, and that the term rigidly designates XYZ *if* the actual world is one where XYZ is the watery substance of our acquaintance, and so on. The analogous holds true for sentences. Here the conditionals connect actual worlds to truth-values.

The 2D Fregean semantics yields a general idea about competence: to be competent with a term or token α is to know its A-intension. To know this is of course to know the role carved out by the descriptive property associated with α, and we can depict what is so known by a set of conditionals of the form 'If w_1 is actual, α applies to the Xs,' 'If w_2 is actual, α applies to the Ys,' and so on. The idea is not that those competent with α share descriptive knowledge associated with α ready to be put into words. The idea is rather that the knowledge is implicit and that competent speakers share an ability: "If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension" (Chalmers and Jackson, 2001, p. 323). Here looms a threat of triviality. The exercise becomes pointless if the information given already presents the actual world in terms of the concept whose extension we try to determine – if you describe the world in terms of grandmothers, figuring out the extension of 'grandmother' hardly poses a challenge

to me. So we need to assume that the information is somehow provided in a neutral fashion (as Chalmers does, see below, §8). Granted that it is, the claim appears plausible for descriptive terms. Tell me who the females and the parents are in some A-world *w*, and I can pick out the grandmothers in *w*, although I still might be hard pressed to put the descriptive property guiding my application into words. According to 2D Fregeans, the same holds good for names and natural kind terms. Consider 'water.' Here it is claimed that "sufficient information about the distribution, behavior, and appearance of clusters of H_2O molecules enables the subject to know that water is H_2O, to know where water is and is not, and so on" (Chalmers and Jackson, 2001, p. 323). This again is a question of being able to apply the term 'water.' It is not a question of being able to put the descriptive property guiding this application into words.

The 2D Fregean semantics also affords a deflationist account of necessities *a posteriori* (see Davies and Humberstone, 1980, p. 18, who also analyze identity statements involving natural kind terms along these lines). Knowledge of A-intensions is assumed to be *a priori*. By knowing the A-intension of 'water,' we know *a priori* that 'water' applies in any A-world *w* to whatever is the watery substance of our acquaintance in *w*. But there are A-worlds where H_2O plays this role, and there are A-worlds where something else does, and no amount of *a priori* reasoning will tell us that the world we happen to live in is of the former variety. This we need to determine empirically. So the A-intension of the *a posteriori* 'Water is H_2O' will come out false at some A-world and is thus contingent. In effect, being *a posteriori* and having a contingent A-intension are two sides of the same coin. Now, 'water' is actuality-dependent in that the term rigidly designates whatever actually fills the role carved out by its A-intension. But the world we happen to live in is one where H_2O plays the role of being the watery substance of our acquaintance. Hence, the C-intension the term 'water' has in our mouths here picks out H_2O in all C-worlds. So the statement 'Water is H_2O' as we actually employ it expresses a necessary C-intension. All other necessities *a posteriori* are accounted for in the same fashion. Any such statement is *a posteriori* in that it has a contingent A-intension, and it is metaphysically necessary in that it has a necessary C-intension. This idea is deflationist in that it makes do without acknowledging propositions that are necessary as well as *a posteriori*.

7 Jackson, or Why Communication Requires a 2D Semantics

Jackson employs his 2D Fregean account to defend general ideas about linguistic meaning. Since his account requires known associations of words with properties (see below), it is not fit to serve also as a theory of mental representation (see Jackson, 2004, p. 275). Jackson's view is rooted in what he deems a folk-theoretic insight into language: Language is a system of representation designed to facilitate the conveying of information. Its key "job is to represent, in a way accessible to actual and potential hearers and readers, how things are according to the speaker or writer" (Jackson, 2005, p. 257). Our words are up to this task because competent speakers resiliently and knowingly associate properties with their words. For example, we associate a specific shape property with 'circle,' and a distinct dispositional property with 'fragile.' These associations need to be resilient in order to allow speakers to rely on them when they employ language to convey putative information. These associations must moreover be known in order for the conveying of putative information to be effective – if you do not know what 'fragile' means, the point of printing that word on a

parcel will be lost on you. Given that the association between a property *F* and a word α in a community is resilient as well as known, *F* is what α as employed in that community semantically contributes to how sentences containing α represent things as being.

The picture sketched assigns sense-like descriptive contents as characterized by (i′)–(iv′) above a key role in semantics. Jackson identifies these descriptive contents with A-intensions in the sense of functions f: $W_A \rightarrow E$ from actual worlds to extensions. This allows him to embrace the general 2D Fregean explanations sketched above. It also puts him in a position to say how A-intensions are fixed: they are put in effect by our conventions linking words to properties, be they direct (as in the association of, say, 'circle' with the respective shape property) or, as will quite often be the case, mediated by our (folk-)theories (as in the association of, say, 'water' with the complex property our folk account of water delineates). On Jackson's view, then, A-intensions are fundamental because they are the basic semantic properties we conventionally assign to our words. Our knowledge of A-intensions moreover comes out *a priori* in a very specific sense. To understand the language I speak is to know which descriptive properties are by convention associated with my words. This suffices to know the A-intensions of my expressions. I do not need to know any further fact. I especially do not need to know what the world that happens to be the actual one is like non-conventionally, and "[w]hat we can know independently of knowing what the actual world is like can properly be called *a priori*" (Jackson, 1998, p. 51). It thus is *a priori* to competent speakers that grandmothers are female for the simple fact that we by convention associate 'grandmother' with the property of *being a female parent of a parent* in a way that does not make this association contingent on any (non-conventional) aspect of the actual world. By contrast, the fact that 'water' designates H_2O is contingent on such an aspect, *viz*. that in the actual world, it is H_2O that plays the role carved out by the descriptive property our conventions assign to 'water.'

Jackson accounts for the rigidity of proper names and natural kind terms by their actuality-dependence (see Jackson, 2004, pp. 261 f.). He also stresses that his view allows competent speakers to have mere implicit knowledge of the associations of words and properties, that the properties concerned might well be causal and egocentric (which is easy to grant given that A-worlds are centered), and that associated properties may vary with the sociolect, rather than the language spoken. Jackson adds a causal-descriptivist account for these terms (see Jackson, 2010; 2005; 2007). The pivotal idea here is that Kripke is right to hold that the reference of proper names is fixed causal-historically, but that competent speakers *know* this. More precisely, a competent speaker knows for any particular name-token α she comes across that its referent "stands at the far end of the information-preserving causal chain […] whose near end is the person making the claim and the token they produce" (Jackson, 2010, p. 141; cf. Jackson, 2007, pp. 20–22). This account explains how we can informationally exploit even names we have never encountered before. It also explains why there is general agreement about what 'Gödel' refers to in Kripke's 'Gödel'-case, which is hard to account for on a pure causal-historic theory (see Jackson, 2007; 2004, p. 273). Still, Jackson's causal-descriptivism explains this at the cost of bestowing rather thin and worryingly schematic senses on proper names in general.

The general positive argument Jackson relies on appears more promising (see Jackson, 2004, pp. 266 ff.; 2005; 1998, p. 40, fn. 16). No one doubts that "[w]ords like 'water,' 'gold,' 'elm,' 'quark,' and so on, are very useful for saying how things are, co-ordinating behavior, transmitting information, and so on" (Jackson, 2004, p. 270). In fact, we can employ these terms to convey putative information to just about any competent hearer (or reader) within our linguistic community, just as we can do with 'fragile' or 'dangerous.' However, for this to

work, Jackson argues that the hearers (or readers) need to knowingly associate properties with the former expressions as well, otherwise we could not account for our success in transmitting information suited to coordinate behavior. Jackson concludes that we in fact knowingly associate properties even with our natural kind terms and proper names. This is precisely the core contention of his 2D Fregeanism.

Jackson tends to present this line of thought as a folk-theoretic insight not yet tainted by philosophical theorizing. This will hardly sway the orthodox Kripkean who anyway subscribes to a revisionist program. She won't find it hard to discard Jackson's truism as a commonsensical semantic prejudice to be revised in the light of Kripke's theory. However, Jackson's line of thought can also be construed as a direct argument aiming to show that we cannot account for our reliable success in linguistic communication suited to coordinate behavior without assuming resiliently associated properties (see Nimtz, 2010, §7). The orthodox Kripkean has a ready reply to this challenge, though. She can argue that her semantics can account for linguistic communication suited to coordinate behavior – provided that we combine it with a Stalnakerian 2D pragmatics.

8 Chalmers's Epistemic Two-Dimensionalism

Judged from the perspective of the philosophy of language, Chalmers's 2D Fregeanism appears more modest than Jackson's. Chalmers's 2D semantics merely aims at tokens rather than types, and he is content to elucidate "utterance content" (Chalmers, 2010, p. 557) rather than shared linguistic meaning. However, Chalmers does not primarily conceive of his project as a contribution to the philosophy of language. His 2D Fregeanism aspires rather to be a general rationalist account of representation and modality establishing A-intensions as a viable kind of content. There are two aspects to this. First, Chalmers's 2D semantics applies to language and mind. He thinks of A-intensions as providing an explanatorily fruitful variety of content both for linguistic tokens and for occurrences of mental states. Second, Chalmers thinks of A-intensions as constitutively tied to modality and to reason (see Chalmers, 2004, pp. 153 f.; Chalmers 2006, pp. 55 f.). A-intensions are constitutively tied to modality in that they concern what is possible and necessary. A-intensions are constitutively tied to reason in that they are defined in terms of what an ideal reasoner can know *a priori* (as we shall presently see). Learning about A-intensions and their conceptual interrelations thus promises to be an *a priori* means to learn about what is possible and necessary. Chalmers welcomes this rationalist upshot. He thinks that *a priori* reflection is indeed a way to learn, *via* insight into A-intensions, about the nature of the mind (see Chalmers, 2010) or the fundamental structure of reality (see Chalmers, 2012). However, Chalmers does not premise his 2D Fregeanism on rationalist ideas. He rather develops a 2D semantics that, if true, would vindicate his rationalism. He goes on to defend this account, reaping a vindication of his rationalism as a consequence.

Chalmers's exposition of his view is comprehensive, complex, and detailed. This is in part owed to Chalmers's pluralism. Chalmers readily acknowledges semantic values other than A-intensions and allows alternative understandings of key concepts. To keep things manageable, I chart just one course through the space of options Chalmers provides.

The key to Chalmers's 2D Fregeanism is an epistemic reading of the A-dimension. We should, he urges, understand A-worlds and A-intensions in terms of ideal *epistemic* possibility. Epistemic possibility specifies what is compatible with what someone does or

can know, rather than what could be the case *simpliciter*. A claim *p* is an epistemic possibility if *p* could be the case *for all we know*. It is an *ideal* (or in Chalmers's terms: *deep*) epistemic possibility if *p* could be the case *for all an ideal reasoner can know a priori*. The simplest epistemic reading straightforwardly identifies A-worlds with maximally specific ideal epistemic possibilities or *scenarios*, as Chalmers calls them. (See Chalmers, 2010, appendix. See Chalmers, 2004, for alternatives.) A-worlds then stand to comprehensive epistemic possibilities just as C-worlds stand to comprehensive metaphysical possibilities. These possibilities differ. Given that 'Hesperus = Phosphorus' is metaphysically necessary, there is no C-world where 'Hesperus = Phosphorus' comes out false. However, there are no *a priori* grounds to rule out that Hesperus ≠ Phosphorus. So there will be an epistemic A-world or scenario where 'Hesperus' denotes something other than 'Phosphorus.'

Chalmers assumes that any scenario is given by a *canonical description* that comprehensively characterizes the scenario in neutral qualitative terms. He relies on canonical descriptions to avoid the triviality of presentation problem mentioned above and to explain what epistemic A-intensions come to. Chalmers identifies A-intensions with functions f: $W_A \rightarrow E$ from epistemic A-worlds to extensions. But what is it for a sentence – say, the sentence 'Venus is bright' – to be true at some epistemically possible A-world? The analogous question for C-worlds is easy to answer. The sentence is true at some C-world *w* just in case *w metaphysically necessitates* it. This holds true just in case the counterfactual conditional '$D_W \,\Box\!\!\rightarrow$ Venus is bright' is true, where D_W is a canonical description of *w*. The same idea applies to A-worlds. Our sentence is true at some epistemic A-world *w* just in case *w epistemically necessitates* it. This holds true just in case the indicative conditional '$D_W \rightarrow$ Venus is bright' is epistemically necessary, that is, *a priori* to an ideal reasoner. Epistemic A-intensions thus "represent the *epistemic dependence* of the extension of our expressions on the state of the world" (Chalmers, 2004, p. 176). So to know the A-intension of a sentence δ is to know *a priori* whether or not δ holds true given that some canonical description *D* truly describes our world.

It is built into Chalmers's epistemic 2D semantics that *apriority* and A-necessity coincide: a sentence token is *a priori* if and only if its A-intension holds true at all A-worlds. (Chalmers sometimes allows unknowable mathematical truths to be an exception to this.) This *core thesis* ties A-intensions to the rational domain. It guarantees that one can learn *a priori* what is A-possible and A-necessary. Now recall the key principle of 2D Fregeanism mentioned above: A-intensions determine C-intensions, either by themselves or together with the actual world. Taken together, these principles explain why, on Chalmers's epistemic account, learning about A-intensions is indeed an *a priori* means to learn about what is non-epistemically possible and necessary. Suppose I have *a priori* determined that 'There is no perfect being' is an epistemic possibility. I thus know that its A-intension holds true in at least one A-world. But the sentence does not contain any terms whose C-intension varies with the actual world. The C-intension of the whole sentence thus cannot differ from its A-intension. I can therefore conclude that the C-intension of 'There is no perfect being' holds true at one C-world, and thereby expresses a metaphysical possibility. Armchair *a priori* insight here directly translates into insight into what is metaphysically possible. When it comes to sentences whose C-intensions do vary with the actual world, things are more complicated. By itself, *a priori* insight into the A-necessity of 'Water = the watery substance of our acquaintance' does not provide insight into a metaphysical necessity. However, we merely need to add the empirical result that the actual watery substance of our acquaintance is H_2O to conclude that 'Water = H_2O' is metaphysically necessary.

Jackson thinks that A-intensions are *a priori* because our associations of terms and properties are purely conventional. Chalmers's rather more robust understanding of the *apriority* of A-intensions rests on his principle of the *scrutability of reference and truth*. Suppose I know a correct canonical description *D* of my world. If reference and truth are scrutable, as Chalmers claims they are, I am then in a position to know, without any further empirical information, what my terms refer to and which of my sentences are true. Chalmers derives this principle of scrutability from the idea that our expressions have a 'normative inferential role.' He explains thus: "[F]or any expression we use, […] given sufficient information about the actual world, certain judgments using the expression will be irrational, and certain other judgments using the expression will be rational" (Chalmers, 2010, p. 555). He goes on to stress that it is "this sort of inferential role that grounds the primary intension of an arbitrary expression (as used by an arbitrary speaker)" (p. 555).

Chalmers argues at length that our expressions do have normative inferential rules (at least for the individual speaker), and that scrutability holds true. His case for the former is positive: he discusses examples where he judges the claim to hold and generalizes from these (see Chalmers and Jackson, 2001). Chalmers's case for the latter is negative: he patiently discusses potential counter-examples to scrutability and finds them all wanting (see Chalmers, 2002b, pp. 174–195). Chalmers clearly believes that his 2D semantics will stand and fall with such overarching principles, and not with its success in explaining humdrum semantic phenomena. This is only partly due to the fact that Chalmers does not intend his 2D theory to be a semantics for a shared natural language. Given his pluralism, Chalmers can easily countenance alternative semantic accounts postulating semantic values other than epistemic A-intensions. Chalmers's semantics is only vulnerable to accounts that, like orthodox Kripkeanism, straightforwardly declare that some of our terms do *not* have epistemic A-intensions, even for the individual speaker.

9 An Upshot, or the State of the Debate

There is little general dispute about the two-dimensional approach within philosophy. Quite the opposite, two-dimensional theorizing is well-established in contemporary philosophy of language. This is in large part due to Kaplan's groundbreaking 2D semantics for indexicals and Stalnaker's somewhat less appreciated 2D pragmatics. There is little dispute about either account. In fact, going two-dimensional by embracing either is seen as a natural and *per se* non-contentious move for theorists working in the Kripkean tradition. In sharp contrast to this, both Jackson's 2D semantics and Chalmers's epistemic two-dimensionalism remain rather controversial, with no agreement on results (see, e.g., Schroeter, 2012; Elliott, McQueen, and Weber, 2013; Speaks, 2014). In part, this may well be owed to the complexity of the issue. To realize their revisionist agenda, 2D Fregeans need to do quite a lot. They need to devise a compelling Fregean semantics that respects Kripke's semantic insights and accommodates Kripke's anti-descriptivist arguments. They also need to convince us that a revision of orthodox Kripkeanism is called for in the first place. The success of any of these tasks is hard to gauge. So the success of any 2D Fregean semantics is likely to always be a matter of some debate. In part, this might well be due to the fact that participants in the debate harbor rather different ideas as to what a semantics should accomplish in the first place. One might well hope, then, that advances in meta-semantics will allow for a more nuanced assessment of 2D Fregean accounts.

Notes

1 I will talk throughout as if the functions and so on that a semantics assigns to expressions are meanings. This is not meant to beg the question against those who, like Kaplan (1977, p. 502), think that these set-theoretic entities merely represent or model meanings.

2 The entities employed in 2D semantics go by different names, and I will indicate who calls what by which name in due course. Still, I need to make a terminological choice, so I follow Jackson (1998; 2004) in my terminology. This is a choice of convenience, not a theoretical commitment.

3 I employ the terms 'semantic' and 'meta-semantic' in their usual meanings. A semantics for certain expressions specifies *which* semantic properties these expressions have. A meta-semantics for certain expressions explains *how* or *why* these expressions came to have those semantic properties in the first place. Kaplan (1989, pp. 573 f.) uses 'meta-semantic' differently. He dubs an account explaining why certain expressions have their semantic properties – a meta-semantics in the standard sense – 'meta-semantic' only if it exclusively enlists non-semantic properties to do the explaining.

4 2D Kripkeans and 2D Fregeans disagree on the fundamentals. However, they can well agree on many specific semantic or pragmatic issues, for example, how to treat indexicals, or how to solve pragmatic puzzles.

5 This makes Kripke's (1971, p. 153) claim that we know conditionals such as 'If this lectern is made of ice, it is necessarily made of ice' by "a priori *philosophical* analysis" (my emphasis) all the more puzzling.

6 $P(\mathrm{D})$ is the power-set of D, that is, the set of all subsets of D.

7 Stalnaker (2014, p. 225) holds that strictly speaking, diagonal propositions involve only uncentered worlds. I ignore this.

8 The following applies to 2D semantics. Things are different with a Stalnaker-style 2D pragmatics.

9 Such a reading would take the phenomenon to be metaphysical, rather than primarily semantical, and undercut the deflationist account of necessities *a posteriori* embraced by 2D Fregeans. See §6.

10 The expression 'dthat(the *F*)' rigidly designates whatever actually satisfies the embedded description 'the *F*.' For example, 'dthat(the second planet from the Sun)' rigidly designates the planet Venus. The claim made in the text holds true on the reading of 'dthat' as an operator only. With 'dthat' read as a term-forming device, the resulting expression is directly referential. See Kaplan (1989, pp. 579–582).

11 Kripke (1980, p. 21) expresses reservations about casting his ideas in terms of 'propositions.' But he has no qualms to talk of 'propositions' in his 1976 John Locke lectures where he evidently conceives of propositions as Russellian (see Kripke, 2013, lecture II). Some (see Soames, 2011, pp. 78 f.) think that a specific view of attitude ascriptions is part and parcel of the view of propositions a Kripkean commits to. I consider this to be controversial, and thus beyond the fourth doctrine as stated. This seems also right for dialectical reasons: Whatever 2D Fregeanism is, it is not *per se* an account of propositions and attitude ascription.

12 Kaplan uses 'content' as a technical term applying only to those contents (in the encompassing sense of the term) that play the role of what-is-said in contexts and are often singular propositions.

References

Åqvist, L. 1973. "Modal logic with subjunctive conditionals and dispositional predicates." *Journal of Philosophical Logic*, 2(1): 1–76.

Burge, T. 1986. "Intellectual norms and foundations of mind." *The Journal of Philosophy*, 83(12): 679–720.

Chalmers, D. 2002a. "On sense and intension." *Philosophical Perspectives*, 16: 135–182.
Chalmers, D. 2002b. "Does conceivability entail possibility?" In *Conceivability and Possibility*, edited by T. S. Gendler and J. Hawthorne, pp. 145–200. Oxford: Clarendon Press.
Chalmers, D. 2004. "Epistemic two-dimensional semantics." *Philosophical Studies*, 118(1–2): 153–226.
Chalmers, D. 2006. "The foundations of two-dimensional semantics." In *Two-Dimensional Semantics*, edited by M. García-Carpintero and J. Macià, pp. 55–140. Oxford: Oxford University Press.
Chalmers, D. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.
Chalmers, D. 2012. *Constructing the World*. Oxford: Oxford University Press.
Chalmers, D., and F. Jackson. 2001. "Conceptual analysis and reductive explanation." *Philosophical Review*, 110(3): 315–360.
Davies, M., and L. Humberstone. 1980. "Two notions of necessity." *Philosophical Studies*, 38(1): 1–30.
Elliott, E., K. McQueen, and C. Weber. 2013. "Epistemic two-dimensionalism and arguments from epistemic misclassification." *Australasian Journal of Philosophy*, 91(2): 375–389.
Evans, G. 1979. "Reference and contingency." *The Monist*, 62(2): 161–189.
García-Carpintero, M., and J. Macià, eds. 2006. *Two-Dimensional Semantics*. Oxford: Oxford University Press.
Hughes, C. 2004. *Kripke: Names, Necessity, Identity*. Oxford: Clarendon Press.
Humberstone, L. 2004. "Two-dimensional adventures." *Philosophical Studies*, 118(1–2): 17–65.
Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Blackwell.
Jackson, F. 2001. "Précis of *From Metaphysics to Ethics*." *Philosophy and Phenomenological Research*, 62(3): 617–624.
Jackson, F. 2004. "Why we need A-intensions." *Philosophical Studies*, 118(1–2): 257–277.
Jackson, F. 2005. "What are proper names for?" In *Experience and Analysis*, edited by J. C. Marek and M. E. Reicher, pp. 257–269. Vienna: Österreichischer Bundesverlag Schulbuch.
Jackson, F. 2007. "Reference and description from the descriptivists' corner." *Philosophical Books*, 48(1): 17–26.
Jackson, F. 2010. *Language, Names, and Information*. Oxford: Blackwell.
Kaplan, D. 1977. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–563. Oxford: Blackwell.
Kaplan, D. 1989. "Afterthoughts." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 565–614. Oxford: Blackwell.
Kripke, S. 1971. "Identity and necessity." In *Identity and Individuation*, edited by M. Munitz, pp. 135–164. New York: New York University Press.
Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.
Kripke, S. 2013. *Reference and Existence*. Oxford: Oxford University Press.
Kuhn, S. T. 2012. "Two-dimensional logic and two-dimensionalism in philosophy." In *The Routledge Companion to the Philosophy of Language*, edited by G. Russell and D. Graff Fara, pp. 624–635. New York and London: Routledge.
Lewis, D. 1980. "Index, context and content." In *Philosophy and Grammar*, edited by S. Kanger and S. Öhman, pp. 79–100. Dordrecht, Netherlands: Reidel.
Nimtz, C. 2010. "Thought experiments as exercises in conceptual analysis." *Grazer Philosophische Studien*, 81: 189–214.
Putnam, H. 1975. *Mind, Language, and Reality*. Cambridge: Cambridge University Press.
Schroeter, L. 2012. "Two-dimensional semantics." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/win2012/entries/two-dimensional-semantics/ (accessed August 27, 2016).
Searle, J. 1958. "Proper names." *Mind*, 67(266): 166–173.
Segerberg, K. 1973. "Two-dimensional modal logic." *Journal of Philosophical Logic*, 2(1): 77–96.

Soames, S. 2005. *Reference and Description: The Case against Two-Dimensionalism*. Princeton: Princeton University Press.

Soames, S. 2011. "Kripke on epistemic and metaphysical possibility: two routes to the necessary a posteriori." In *Saul Kripke*, edited by A. Berger, pp. 78–99. Cambridge: Cambridge University Press.

Speaks, J. 2014. "No easy argument for two-dimensionalism." *Australasian Journal of Philosophy*, 92(4): 775–781.

Stalnaker, R. 1978. "Assertion." *Syntax and Semantics*, 9: 315–332.

Stalnaker, R. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stalnaker, R. 2004. "Assertion revisited: on the interpretation of two-dimensional modal semantics." *Philosophical Studies*, 118(1–2): 299–322.

Stalnaker, R. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

Stalnaker, R. 2014. *Context*. Oxford: Oxford University Press.

van Fraassen, B. C. 1977. "The only necessity is verbal necessity." *Journal of Philosophy*, 74(2): 71–85.

38

The Semantics and Pragmatics of Indexicals

JOHN PERRY

1 Introduction

The term 'indexical' comes into the philosophy of language from Charles Sanders Peirce's use of the term 'index.'[1] Here is an explanation of Peirce's threefold division of signs: icon, index, and symbol:

> Signs are icons, indices (also called "semes"), or symbols … accordingly as they derive their significance from resemblance to their objects, a real relation (for example, of causation) with their objects, or are connected only by convention to their objects, respectively. (Burch, 2014)

Suppose I am talking face-to-face with Elwood Fritchey. There is a completely arbitrary convention that allows me to refer to him as 'Elwood.' Although in this case he is standing before me – a "real" relation – that has nothing to do with the fact that I refer to him when I say "Elwood." I will be able to refer to him by using 'Elwood' after he has gone. The work of securing reference is done by the convention, and does not depend on any further connection between the speaker and Elwood. So 'Elwood,' it seems, is a symbol, connected to Mr Fritchey only by convention.[2]

One of Peirce's central examples of indexicals was smoke, which is a sign of fire. Here the real relation is causation: fire causes smoke. No convention is involved, and no language. Smoke is a natural sign of fire, not a conventional one.

But the term 'indexical' is not much used for natural signs in contemporary philosophy of language. It is used for words like 'I,' 'you,' 'here,' and 'now,' which clearly have conventional meanings. However, their conventional meanings do not determine the reference of uses of such expression all by themselves. I can use 'you' to refer to Elwood, given its conventional meaning, because Elwood is the person I am talking to. This is a real relation between Elwood and Me, involving causation and perception. This kind of conventional indexicality

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

is the phenomenon for which "indexicality" is used in contemporary philosophy of language, and will be our topic here.

Paradigm indexicals include pronouns such as 'I' and 'you,' as well as words like 'here,' 'now,' 'today,' 'tomorrow,' and 'yesterday' that occur as both nouns and adverbs. We'll look at how such paradigms work, then look at less paradigmatic examples, and eventually try to arrive at plausible definitions of *indexical* and *indexicality*.

When I say "I," I refer to myself. When you say "I," you refer to yourself. We use the same word, with the same meaning, but we refer to different persons. This is a characteristic of indexicals; the meaning of the expression does not wholly determine its *reference* on a given occasion of use, or as I shall say, the reference of a given *utterance* of the expression.[3] The meaning of an indexical determines its reference only in combination with the *context*, that is, in combination with facts about the particular utterance. In the case of 'I,' the relevant fact is who the speaker is. The meaning of the expression 'I' is captured by the rule: an utterance of 'I' refers to the speaker of that very utterance. Similarly, for our other paradigms:

An utterance *u* of 'you' refers to the person the speaker of *u* is addressing;
An utterance *u* of 'now' refers to the time of *u*;
An utterance *u* of 'today' refers to the day that includes the time of *u*;
An utterance *u* of 'tomorrow' refers to the day after the day that includes the time of *u*;
An utterance *u* of 'yesterday' refers to the day before the day that includes the time of *u*.

Generalizing from these examples, we get a simple theory of indexicals:

If α is an indexical, the *meaning* of α is a function from utterances to objects that is conventionally assigned to α. The referent of an utterance *u* of α is the value of that function with *u* as argument.

In the cases of 'I,' 'here,' and 'now' the relation arguably provides a *total* function; every utterance has a speaker, location, and time.[4] But in many cases we have only *partial* functions. I think I hear a knock at the door, and say "You can come in." But it was just an odd noise; there is no one at the door; I'm not addressing anyone; my utterance of 'you' doesn't refer to anyone. More subtly, I may start my utterance of 'today' just before midnight, and end it just after. There is no one day that includes the time of my utterance. Or, more dire, perhaps today is the end of days. Then my use of 'tomorrow' doesn't refer to anything. And working on this chapter is rather futile.

This simple theory will need to be modified. But it will suffice to make some initial points, distinctions, and contrasts with other approaches.

2 Approaches

It's not very difficult to see how paradigm indexicals work. The difficulties and controversies have more to do with how to fit accounts of indexicals into more general theories of meaning, cognitive significance, and pragmatics.

Eliminative theories treat indexicals as short-cuts for descriptions that the speaker has in mind. Indexicals play no essential role in language and communication and can be eliminated for serious purposes. Perhaps I think of myself as Stanford's oldest living non-Norwegian philosopher. Then perhaps we would translate my utterance "I moved to The

Bay Area in 1974" as "Stanford's oldest living non-Norwegian philosopher moved to The Bay Area in 1974."

This approach is not very plausible. I haven't always thought of myself in this way. But I seem to say the same thing with my utterance of "I moved to the Bay Area in 1974" as I would have said using the same sentence five or ten years ago. But then I had older non-Norwegian colleagues, and couldn't have identified myself in this way. The proposed translation would have been false. Further, if I now explicitly say, "Stanford's oldest non-Norwegian philosopher moved to The Bay Area in 1974," the self-knowledge expressed by I saying "*I* moved to The Bay Area in 1974" would be missing. I might know that Stanford's oldest non-Norwegian philosopher moved to the Bay Area in 1974, without realizing that *I* did so. We might, following Russell, try to replace "I" with something like "the person having *this* thought now." But we would have simply traded one indexical, "I," for two, "this thought" and "now." It's very hard to come up with any plausible candidate to translate "I" that does not employ an indexical; we'll look at the reasons for this below.

The eliminative approach has seldom been defended; it is important only as what philosophers have perhaps had in mind, who ignored the phenomenon of indexicality.

A second approach I call *reductive*. Reichenbach's "token-reflexive" theory is an example (Reichenbach, 1947). Reichenbach reduced all indexicality to one expression, 'this*,' which means roughly, "this very token." So 'I' means "the speaker of this*," 'you' means "the addressee of the speaker of this*" and so forth. Alternatively, one might suppose, following Castañeda (1999, esp. chapters 1 and 2), that all indexicals can be defined in terms of 'I' and 'now.' 'You,' for example, means "the person I am addressing now" and 'here' means "the location I am now in." Such reductive approaches allow translation into a formal language based on the predicate calculus, which in its standard form has no provision for context relativity, with a minimal amount of change.

From the last part of the last century, however, the idea that translation into the predicate calculus takes care of semantics has waned, in favor of the idea that we should develop semantics for languages that can treat problematic expressions on their own terms. Richer formal languages, in particular modal and intensional logics, were developed. And context-sensitive expressions, like indexicals, came to be seen as rather central parts of language, rather that outliers to be ignored or kept to a minimum. In David Kaplan's theory (Kaplan, 1989), the most influential account of indexicality, indexicals are full-fledged parts of an intensional language and are given a semantics directly, rather than by translation.

Kaplan eschews utterances in his formal theory, in favor of *sentences in context*. A context, in Kaplan's technical sense, is a sequence of an agent, location, time, and world, where the agent is in the location at the time in the world. Expressions have *characters*, which are functions from contexts to *contents*. Contents are propositions and propositional constituents. So the character of 'I' is a function from contexts to the agent of the context. The character of "I moved to The Bay Area in 1974" is a function from a context to the proposition that the agent of the context moved to The Bay Area in 1974, that is, in the case of my utterance, the proposition that JP moved to The Bay Area in 1974. Thus suppose Michael Bratman and I each say truly, "I moved to the Bay Area in 1974." Our utterances have different *contents*, they express different *propositions*, both true as it happens.

A word about propositions. Propositions are abstract objects that encode truth-conditions, and are usually taken to be what 'that'-clauses stand for in statements like "Michael said that he was born in the Bay Area in 1974." *Qualitative propositions* rely only on properties and relations to encode truth-conditions: such propositions specify the patterns of instantiation

and co-instantiation needed for truth. *Singular propositions* allow particular individuals to do the instantiating and co-instantiating. Thus, if we take propositions to be sets of possible worlds, the singular proposition that Michael Bratman lives in the Bay Area would be the set of worlds in which Bratman himself instantiates the property of living in the Bay Area, whatever he is called, and whatever his most distinctive properties might be, in those worlds. It is true because the actual world is one of these worlds. Another conception of propositions is as sequences of properties and objects, in this case the sequence: <*living in the Bay Area*, Bratman>. The proposition is true because Bratman has that property. Many of us were convinced of the need to recognize singular propositions, conceived of in one way or another, by considerations about *rigid designation* and *direct reference* which I discuss below in §7. Others remain skeptical.[5]

Given that Bratman and I express singular propositions, they seem to be the same singular propositions we could have expressed by saying, respectively, "John Perry moved to The Bay Area in 1974," and "Michael Bratman moved to The Bay Area in 1974." The quite different ways of referring to ourselves, with 'I' or with our names, seem to be lost at the level of the proposition expressed, an issue I deal with below in §4.

3 The Semantics of Indexicals

The theory I develop here is inspired by Kaplan's approach, but has evolved in ways that incorporate significant differences (see Perry, 2011; forthcoming b). Most important is the explicit inclusion of utterances in the theory. To understand the pragmatics and cognitive significance of indexicals, we need to understand how the properties of utterances – particular episodes that are part of the causal realm – interact with other sorts of episodes, in particular cognitive episodes. For this purpose it is useful to have utterances in our theory.

The difference is a bit subtle. In Kaplan's formal theory, we have "sentences in contexts" rather than utterances. Suppose I say, at some time in the afternoon on November 29, here in Palo Alto, "The San Francisco Warriors play tomorrow." On Kaplan's theory we have the context: <JP, Palo Alto, November 29>. This context, plus the sentence I used, give us a content: that the Warriors play on November 30.

Contexts are clearly inspired by the nature of utterances. Utterances involve a speaker uttering an expression at a locations and time. From the point of view of developing a logic of indexicals, Kaplan finds it useful to abstract from the *uttering*. Instead of an utterance we have a context – agent, location, time, and world – and an expression. The combination of expression and context captures key facts about utterances, and so can be used to model them. But we can assign contents to the combination of context and expression even if there is no utterance – where the *agent* of the context isn't a *speaker* at the time and location in the worlds under consideration.

On this approach, the relation between the utterances and the elements of context are not dealt with in the theory. We *model* utterances as sentences in contexts, and then the theory takes over.

But for certain purposes, it is helpful to have utterances within the theory. Consider a belief an utterance that expresses it. Here we have a causal relation between a mental state and an intentional act and a content relation between the two. A theory that seeks to understand this complex relation, involving both cause and content, needs to be able to explicitly recognize the states and events that stand in the complex relations.

Suppose, on the 28th of the month, I leave the message, "The Warriors play tomorrow" on my daughter's answering machine. She doesn't check her answering machine, however, until the next day, the 29th. She interprets the recorded message she hears as meaningful, because she takes it to be the result of an intentional utterance. To correctly interpret it, she has to figure out *when* the utterance was made. Perhaps she gets this wrong, and thinks my call occurred on the same day she heard it and so wrongly takes it that the Warriors play on the 30th, rather than the 29th, as I intended to convey. With written communication, and recorded communication, the token produced by an utterance is not always perceived at the time of the utterance. The hearer or reader needs to make inferences to figure out the time of the utterance (or the speaker, or the location) to correctly interpret indexicals. This reasoning is based on the nature of utterances, their effects, and so on. Thus to understand my daughter's reasoning, and why the message I left was perhaps a bit inept, it is useful, at least, to have utterances in our theory.

Kaplan's *characters* are functions from contexts to referents associated with expressions. A character is a function from contexts to appropriate referents, that turns on the elements of context. For example, 'tomorrow' at *c* refers to the day after the day that includes the time of *c*.

What I call *roles* are similar to characters. Expressions provide a function from *utterances* to referents, that in the case of indexicals, turn on the speaker, location, and time of the utterance. So, an utterance *u* of 'tomorrow' refers to the day after the day that includes the time of *u*.

Metaphysically, the difference is a bit bigger than it sounds. Utterances are concrete events with many properties, including causes and effects. So, having utterances in the theory gives us a link to the mental events and states that lead to utterances, and their real and intended effects – the stuff of pragmatics. A context is an abstract object, a set-theoretical sequence of agent, location, time (and world), with no causes and no effects.

But, in practice, the difference is less than it sounds. Most philosophers use Kaplan's theory to think about utterances, rather than to prove theorems in logic. My approach makes explicit what is usually implicit in such applications of his account.

I assume that an utterance is an event or episode that lasts for a (relatively short) interval of time, and has a speaker who makes use of an expression, and a location. So, for each utterance *u* we have an expression E_u, a speaker, S_u, a time, T_u, and a location L_u. The location of the utterance is the location where the speaker uses the expression at the time.

Let *R-α* be the role associated with expression α. *R-α* is a (possibly partial) function from an utterance of α to its referent. The function will depend on some combination of the speaker, time, and location of the utterance. Since the location is determined by the speaker and time, we can set it aside when not relevant. So we have, as our preliminary analysis,

> The reference of an utterance *u* of an indexical α is the value of the associated role *R-α* with the speaker and time of *u* as arguments:
> The reference of an utterance *u* of α = *R*-α(S_u, T_u).

4 Cognitive Significance and Pragmatics

Suppose I say, "You look tired now" to Dikran at a time I'll call '**t**.' Call this utterance **u**. We can give three different, but consistent, truth-conditions for **u**:

E-conditions: **u** is true iff
$\exists x, t$ such that

(i) $x = R\text{-you}(S_u, T_u)$
(ii) $t = R\text{-now}(S_u, T_u)$
(iii) x looks tired at t.

I call these "E-conditions" because they are conditions on the utterance, that is the *Episode*.

S-conditions: Given that JP is the speaker of **u** and **t** is the time of **u**, **u** is true iff $\exists x, t$ such that

(i) $x = R\text{-you}(\text{JP}, \mathbf{t})$
(ii) $t = R\text{-now}(\text{JP}, \mathbf{t})$
(iii) x looks tired at t.

I call these "S-conditions" because they are conditions on the *Speaker* (and time) of the episode.

O-conditions: Given that JP is the speaker of **u**, **t** is the time of **u**, Dikran = R-you(JP, **t**) and **t** = R-now(JP, **t**), **u** is true iff:

Dikran looks tired at **t**.

I call these "O-conditions" because they put conditions on the *Objects* that the speaker is referring to and talking about.

O-conditions provide singular propositions, and correspond to Kaplan's *contents*. We would usually take what I said to be the proposition that Dikran looks tired at **t.** If Dikran were to say, "Yes, I look tired now," we would take him to have said the same thing that I did. And if someone were to say, "JP said that Dikran looked tired at **t**," we would take that as an accurate report of what I said. Dikran is referred to in three different ways, but the O-conditions, and the singular propositions they determine, are the same.

The level of O-conditions gets at *what else* has to be the case, for an utterance *u* to be true, given the reference-determining facts. This level is important because the reference-determining facts are often common knowledge. The O-conditions get at the additional or *incremental* information provided by the utterance. This importance is reflected in our practice of treating the three utterances as "saying the same thing."

O-conditions obscure differences among the three utterances, however, which are important in understanding the cognitive significance of utterances and their *pragmatics* – how they are used in communication. S-conditions and E-conditions are useful at keeping track of these differences; explicitly considering them is a second departure from Kaplan's approach.

Suppose that on December 5 I ask Dikran when the next meeting of our Questions Group is. As it happens, it is scheduled to meet that very day. He can answer me truly in two different ways:

"The next meeting is today"
"The next meeting is December 5"

Since "today," uttered on December 5, and "December 5," both refer to the same day, we *can* say that the two remarks would have the same truth-conditions: that the next meeting is December 5.

But the two remarks have different *cognitive significance*. That is, the speaker's beliefs that motivate the remarks, and the beliefs the speaker can expect the credulous hearer to acquire, may differ. What I would learn from the first differs from what I would learn from the second, at least if, as usual, I didn't know what day it was. In that case, the first would be more helpful. And Dikran might have been quite sure that the first remark was true, even if he didn't know what day it was, and so was not in a position to make the second remark.

We can get at the difference at the level of S-conditions. The first remark is true if the meeting is to occur on the same day that Dikran and I are conversing. So it provides me with the information I need for timely preparation for the meeting, even if I don't know what the date is. The second provides me with this information only if I already know that today is December 5. On the other hand, if I wanted to write down the day of my meeting on my calendar, Dikran's first utterance wouldn't give me all the information needed, while his second would.

The roles that are exploited by referring expressions are often associated with what I call *epistemic* and *pragmatic* methods (see Perry, 2011). There are ways of finding out more about the day referred to as 'today' that one can only employ on that day: look around and see what is happening, remember what has happened since the day started, and wait and see what else happens before the day ends. There is a way of finding out about the day referred to as 'December 5.' Look on your calendar at the cell with that date as a label, and see what happened, or is expected to happen, on that day.

Similarly, there are ways of acting, or developing plans for action, that are appropriate for days referred to as 'today' and 'tomorrow.' Dikran's first remark tells me I'd better start getting ready for the meeting. With his second remark, the appropriate action seems to be to enter the meeting on my calendar in the appropriate cell, and then figure out whether December 5 is today, or tomorrow, or next week, and develop a plan for getting ready.

As the example suggests, the way we use expressions, their pragmatics, is responsive to the utterance-relative roles associated with the expressions and the epistemic and pragmatic methods connected with those roles. If Dikran wants me to get ready, and is aware of my usual ignorance of what day it is, he would be remiss if he made the second remark rather than the first.

For another example, suppose it's January 16, 2015, and I want to tell Dikran that it's my 72nd birthday. I could say,

> "I turn 72 today"
> or
> "John Perry turns 72 today"

In each case, the way I refer to the objects I am talking about – myself and January 16, 2015 – contains information about them. Dikran knows that my utterance is true if the person I am referring to is 72 on the day I refer to. By referring to this person as 'I,' I provide a second mode of presentation or channel of information for this person. Dikran will realize:

> The person referred to by the utterance I am hearing = the speaker of that utterance = the person sitting across from me, and with whom I am talking

Suppose I'm not talking to Dikran, who knows me and knows my name, but a stranger, who asks me why I look so morose. "I am 72 years old today," does fine as a helpful explanation. "John Perry turns 72 on January 16, 2015," or even "John Perry turns 72 today" would just be puzzling. I don't look a day over, say, 68, so it might not occur to the stranger that I am talking about myself. The natural response might be, "Well, who is John Perry, and why feel sad that he's getting old?"

Now if I said, "John Perry turns 72 today," to Dikran, he could respond appropriately since he knows my name. But my statement would be a bit odd. This is because I am putting an unnecessary cognitive demand on Dikran. To respond appropriately, he needs to recognize me and remember my name. No doubt he would, but why require the extra steps, when the indexical formulation gives him all the information required?

Charles DeGaulle had the habit of referring to himself in the third person, as does Bob Dole, Bill Clinton's opponent in the 1996 U.S. presidential election. At a political rally in Philadelphia, while campaigning against Clinton, Dole said, probably alluding to Clinton's treatment of White House interns,

> If something happened along the route and you had to leave your children with Bob Dole or Bill Clinton, I think you would probably leave them with Bob Dole.

New York Times columnist Ellen Goodman's commented,

> I am not at all sure that I'd want to leave my children with someone who talks about himself in the third person. [6]

Goodman found Dole's utterance odd, probably rather pretentious. Dole's way of putting his point assumed that everyone who heard him knew that 'Bob Dole' was his name. Even though the assumption was doubtless true or nearly so, it still sounds pretentious and self-important. And, even given my well-know humility, Dikran might have the same reaction if I started referring to myself in the third person in conversations with him.

On the other hand, suppose I call Dikran on the telephone. "Who is this," he asks. "It's me," I reply. This may be unhelpful, if he doesn't recognize my voice, and in any case will strike him as annoying and a bit pretentious. To carry on an intelligent conversation, he needs to know more about the person on the other end of the line than merely that he is the person on the other end of the line. He needs access to the information about me he associates with the name 'John Perry,' and to be helpful I should refer to myself in a way that supplies it.

The suggestion, then, is that one important aspect of the pragmatics of indexicals is that referring to something indexically can open new "channels of information" about the things referred to, and indexicals are appropriate when these channels are relevant, and inappropriate when they are not. To understand how this works, we need to consider S-conditions and E-conditions and not just O-conditions. This idea is developed in my book with Kepa Korta, *Critical Pragmatics* (Korta and Perry, 2011).

5 Two Distinctions

Things are often a bit more complicated than envisaged by our preliminary theory, and these complications provide some ways of categorizing indexicals. I classify indexicals by two criteria, whether they are *automatic* or *discretionary*, and whether their reference

Table 38.1 Types of indexicals.

	Narrow	*Wide*
Automatic	'I,' 'today,' 'yesterday,' 'tomorrow'	'yea'
Discretionary	'here,' 'there,' 'now,' 'then'	'this,' 'that,' 'this building,' 'that man,' 'he,' 'she'

depends on narrower or wider circumstances of utterance. Table 38.1 includes the paradigms and some others. The reference of discretionary indexicals depends on the speaker's intentions in ways not yet provided for. The utterance-relative roles associated with them *constrain*, but do not completely determine, the referent.

Of course, with all expressions the intention to use them with their ordinary meanings is relevant. However, it seems that with a word like 'I,' no further intention is relevant to determining the referent. If I am speaking English, 'I' refers to me when I use it. The indexicals 'now' and 'here' seem at first glance as automatic as 'I.' But with 'now' there is a question of how long an interval of time is counted as the present moment; with 'here' there is a question of how much of the surrounding territory is included. The speaker has some discretion in what an utterance of these expressions stands for.

Suppose I say, "Now that the Supreme Court has decided that corporations have the right of free speech, campaign finance reform is hopeless." The period of time I refer to includes the moment of my utterance, but is not limited to that; it is a period that begins more or less when the court decided the Citizens United case, and will continue until their decision is changed, or perhaps reversed by amendment. In contrast, if my wife asks when the TV show *Downton Abbey* starts, in order to plan her evening, and I say "now," I'll have a much shorter interval in mind. Utterances of 'here' behave similarly.

'Today' seems automatic, and that's how I've listed it. But to deal with certain cases, we'd have to demote it to discretionary. Suppose I send an email to an editor in Paris concerning an overdue manuscript. "You will get it today," I write. To fulfill my promise, I have to send the manuscript on a day that includes the time of my message, that is, a 24-hour period from midnight to midnight that includes the time of the utterance. But midnight Paris time, or midnight California time? The editor hopes I mean the former, and a more diligent author might have had that intention. I probably would mean "today" California-time, giving myself a few extra hours, and count on the editor realizing this too late to complain.

With discretionary indexicals, the associated utterance-relative roles *constrain* but do not fully determine reference. We need to alter the rules in a significant way. If 'now' can be used to refer to indefinitely many stretches of time including the moment of utterance, something else needs to account for why a speaker refers to one of the eligible stretches rather than another, and this burden seems to be carried by the speaker's intention.

For example:

> An utterance *u* of 'now' refers to stretch of time *t*, *iff t* includes T_u and the speaker intends to refer to *t*.
> An utterance *u* of 'here' refers to a region *r* iff *r* includes L_u and the speaker intends to refer to *r*.

Ordinarily such intentions will be rather vague. If I say, "it is snowing here," do I mean to refer to my neighborhood, or Palo Alto, or the Bay Area? It's possible that I don't have any

intention that determines among these alternatives, although it may be clear enough that I don't mean to include all of California.

Some indexicals are sensitive only to the *narrow* context, that is, the speaker, time, and location of the utterance. Others are sensitive to elements of the *wider* context. My favorite example is 'yea.' "The fish I caught was yea long," means that it was as long as the distance between my outstretched hands as I speak. This indexical, perhaps not that common outside of Nebraska, requires a demonstration on the part of the speaker, something in addition to the speaker, time, and location, so it is in the wide column of our chart. But, at least as I construe it, the speaker has no option but to refer to the distance demonstrated, so it is in the automatic row.

Kaplan included demonstratives among indexicals. With a simple demonstrative like 'that,' the speaker's intention, sometimes indicated with a demonstration, is relevant. The same speaker at the same time and place might say "that" to refer to any salient object, perhaps something clearly in view, or perhaps something already mentioned in discourse.

Pronouns like 'he' clearly have an indexical use; it seem synonomous with 'that man,' or nearly so. Dikran and I are talking about Obama; as we speak a man walks by wearing a motorcycle helmet. "He is ready for anything," I say. The wider circumstances determine that both Obama and the fellow walking by are salient, but not which of these salient persons I refer to; that depends on my intentions. So 'he,' like 'that man' goes is the cell for indexicals that are both wide and discretionary.

6 Indexical and Undexical Uses

Dikran and I are planning our evening. "Let's go to a local bar," he suggests. The adjective 'local' seems function indexically, amounting to pretty much the same as 'near here.' It seems that:

> An utterance u of 'local' refers to the region in the neighborhood of L_u.

But now suppose I tell Dikran of an upcoming trip to Ireland. "Everywhere you go, be sure to have a beer at the a local bar," he says. 'Local' isn't functioning indexically here. But it seems intuitively to have the same meaning as in the first example. What is going on?

'Local' is associated with the function, *the neighborhood of*, where the argument for the function is a location of the sort an individual occupies at a given time, and the value is a wider region that constitutes the neighborhood of that location. In Dikran's first remark, the argument for this relation is simply L_u, the location of his utterance u: we comply with his suggestion if we go to a bar that is in the neighborhood of the place where he makes his remark. But in the second remark, the argument is provided by the quantifier, 'everywhere.' He means, roughly,

> While you are in Ireland, if you are in a location l, have a beer at a bar in the neighborhood of l.

In the case of Dikran's second utterance, his use of 'local' doesn't refer to a single region determined by the location of his utterance. It functions as a complex variable. It takes the locations quantified over by "everywhere you go" as arguments, and returns the local neighborhood for each.

Finally, suppose Dikran says, one afternoon on campus, "Remember our trip to Copperopolis? There was a local bar that was quite rowdy." Here 'local' refers to the neighborhood of Copperopolis; the argument for the function *R-Local* is Copperopolis, not Stanford, the location of his utterance.

I shall say that a word like 'local' that *can* take the location (time, or speaker) of the utterance as the argument for the associated role has *indexical uses* and exhibits *indexicality*. If the same word can also pick up the argument in other ways, it has *undexical uses* (see Perry, forthcoming a).

Some indexicals seem to have undexical uses. It is natural to think of 'past,' 'present,' and 'future' as indexicals, referring to stretches of times before, at the same time as, and after the time of the utterance. But J. M. E. McTaggart said, early in the twentieth century, of Queen Anne's death, which occurred in 1714:

> It began by being a future event. It became every moment an event in the nearer future. At last it was present. Then it became past, and will always remain so, though every moment it becomes further and further past. (McTaggart, 1908, p. 460)

McTaggart clearly doesn't mean that if you go back far enough, Queen Anne's death occurred subsequent to his writing that paragraph in the early 1900s. He is using 'future' and 'past' undexically.

One might suppose this is an abuse of language permitted only to philosophers. But here is a remark from an 2014 internet discussion of concussions:

> After her concussion in 2005, my 15-year-old daughter was very worried about whether she could play soccer again in the near future.

The writer doesn't seem to have a philosophical agenda, but is clearly using 'future' for events subsequent to 2005, not 2014.

A familiar piece of parental advice goes,

> Never put off until tomorrow what you can do today.

Here the functions associated with 'today' and 'tomorrow' seem to pick up arguments from the quantifier 'never.' If my mother told me this on, say, January 20, 1948, she wouldn't mean merely to suggest that I do things on January 20 and not leave them until January 21, but that *whenever*, throughout the rest of my life, I find myself with a task to do on a certain day, I should do it that day rather than put it off until the next one.

David Kaplan's theory of indexicals doesn't allow for such undexical uses of indexicals; he calls them "monsters." Handling them within his system would seem to require quantifying over contexts, which he doesn't allow.

The present system can be more flexible, however, because we have utterances as part of the theory. Contexts as such play no official role in the theory; the work of being input arguments for the roles associated with indexicals falls to the objects that play the roles Speaker-of, Time-of, and Location-of relative to the utterance in question. We can take the utterance to be the *default* source of the arguments, while allowing for other uses, without quantifiying over contexts. In the case of my mother's advice, the variable assignment associated with the quantifier 'never' supplies the arguments. In the case of the soccer parent's remark, it is the time of the concussion in 2005.

So, for example, we can suppose that the default for 'tomorrow' is an indexical use, so an utterance *u* of 'tomorrow' refers to *R-Tom*(S_u, T_u). But the arguments for *R-Tom* can be supplied anaphorically, as with the remark by the parent of the injured soccer player. Or it can be supplied by a quantifer, as with my mother's advice.

Among paradigm indexicals, the default seems especially weak with temporal indexicals. 'Here' and 'there' also seem to permit undexical uses, as in "You never seem content to be here. You always want to be there." But it's hard to think of a convincing case with 'I.' I'll leave that as an excerise for the reader. If you have a good example, send me an email.

There are clear non-indexical uses of 'here,' as when one points to a spot on a map, and says, "We'll start our trip here." We might postulate an ambiguity, but it seems we might also treat such cases as undexical uses, without introducing a new meaning. The meaning in such uses is still *R-Here*. In the undexical uses, we refer to the region close to the location represented by the demonstrated spot on the map, rather than the location of the utterance itself.

I'll use the term *indexicality* for the expressions that are associated with a reference-fixing function that *can* take the speaker, time, and location of their utterance as arguments. I'll use *indexicals* for expressions that exhibit indexicality, and where the indexical use is the only use permitted, or the clear default in normal conversation. So, in my terminology, 'local' exhibits indexicality, but is not an indexical.

7 Tokens and Technology

I distinguish between utterances and tokens. In my use, utterances are events; they belong to the same general category as other episodes, like perceptions and thoughts, that give rise to them. More specifically they are acts. Tokens are things produced by utterances, of various sorts: sounds that travel from speaker to hearer, or, given the technology of writing, marks on paper.

Primordial conversation was face to face. This meant the time of utterance and the time of perception of the uttered token were, for all practical purposes, the same. The location of the speaker and the location of the hearer, on the other hand, could be relevantly different in some cases. There was a natural dichotomy between 'here' and 'there,' the first for the speaker's position, the second for the hearer's. Not so with 'now.' The contrast between 'now' and 'then' would not be between speaker's time and hearer's time, but time of utterance and some other salient time. Speakers and hearers shared their nows, but not always their heres, to put it undexically. Shouting and smoke signals allow communication at considerable distance, and the telephone takes that further – but the time of utterance and time of perception remain the same. With written language, however, the time of the utterance that produces a token and the time or times at which the token is perceived can be distant from one another.

The practice of copying gives us an important relation among tokens. There is the original, directly produced by the utterance, copies of it, copies of those copies, and so on. Certain properties persist along this relation. The token is evidence for the existence and nature of the utterance that produced it, and so evidence of the intentions of the speaker (by which I will include the writer except when noted). And the copies of that token also provide such evidence, although perhaps not with as much certainty. Other properties do not persist. Sherlock Holmes can tell from a note whether the writer was right- or

left-handed, but he can't tell that from a copy someone else made of the note. For certain legal purposes, the original is required.[7]

In Kaplan's theory 'I am here now' is a theorem in the logic of demonstratives. It does not express a necessary truth, true in every possible world, but it is true in every context. This is because Kaplan required the speaker of a context to be at the location of the context at the time of the context in the world of the context. In our utterance-based theory, every utterance of "I am here now" is true because of the nature of utterances: the speaker of the utterance is at the location of the utterance at the time of the utterance.

Kaplan devoloped his theory in the 1970s. By the 1980s, answering machines had proliferated, and "I am not here now" became an oft-heard and easily understood and believed message. Is there something wrong with Kaplan's theory? Or something wrong with the 1980s?

One might suppose that it has become conventional to refer to the time of listening to a recording with 'now.' "I am not here now" means more or less that the speaker is not at the place where the utterance was recorded at the time the listener hears the resulting token.

However, messages like

> I've got to leave now. I won't be here when you hear this recording. I'll try to call you back later.

in which the use of 'now' refers to the time of utterance, are permissible and intelligible. It seems to be the speaker's choice, whether to use 'now' for the time the recording is made or the time it is heard.

Stephano Predelli (1998) notes that the phenomenon in question pre-dates answering machines. Only the technology of leaving notes is required. Predelli's character Jones has to flee unexpectedly; he leaves his wife a note:

> [P] As you can see, I am not at home now. If you hurry, you'll catch the evening flight to Los Cabos. Meet me in six hours at the Hotel Cabo Real.

The token produced in this case is the note. Jones has a plan, that his wife see the note when she returns at 5 p.m. He uses 'now' to refer to that time, not to the time when he writes the note.

Again, we cannot simply suppose that there is a convention with notes to use 'now' for the time of token-perception. Jones could have written:

> [P′] I'm leaving now for Los Cabos. I'll have been gone for a long time by the time you read this when you get home. If you hurry, you'll catch the evening flight to Los Cabos, and can meet me by 11.

If Jones had written P′, the use of 'now' would have referred to the time he left the note.

There is an interesting and insightful literature on how to think about this. Here I will simply consider whether the idea of undexical uses of indexicals helps. The proposal is that with [P], the speaker uses 'now' undexically. The argument for *R*-now is the time of the perception of the note, not the time of its production. Jones thinks of this time descriptively: the time during which his wife will read the note. His wife thinks of this time as 'now,' used indexically. In [P′], however, 'now' is used in the default way; it is the time Jones thinks of as "now," used indexically, and his wife thinks of as "when my husband wrote the note."[8]

8 Demonstratives

I'll call 'this' and 'that' *simple demonstratives*. Phrases like 'this man' or 'that building,' which add a general term, I'll call *augmented demonstratives*. Jeff King (King, 2001) has argued that what he calls "complex demonstratives," phrases like 'that student who scored one hundred on the exam,' should be treated as quantifiers rather than as indexicals. King's arguments are subtle and complex, and I won't deal with them here, but focus on simple and augmented demonstratives.

It seems that a use of 'that building' will be used to refer to a building that a speaker has in mind and intends to refer to. Typically, at least, the speaker will expect his hearer or hearers to be able to pick out, one way or another, the building in question. I'll say the referent of a demonstrative expression will, in paradigm cases, be *salient* to both speaker and hearer. For example, Dikran and I are walking through the new Engineering Quad at Stanford, and I say, "I like the arches on that building." I might be referring to a building that we are both looking at; it is perceptually salient. If not, I may accompany my remark with a demonstration, pointing at the building I have in mind. Or, if there is only one building with arches, I might count on Dikran using common sense and the charitable assumption that what I say makes sense, and to look around until he sees it.

Back at the office, Dikran might say, "The arches on the building aren't much like the ones in the original quad." The building is not peceptually salient, but is *conversationally* salient; he is referring to the building we were talking about earlier. And there are many other ways an object might be salient, or easily become salient, in a way that makes it suitable to be referred to demonstratively. There is a "degree of freedom" with demonstratives, the expressions on our chart in the cell for indexicals that are both wide and discretionary, that is a bit different than with our paradigm indexicals; the speaker's intentions determine the type of salience relation (*Sal*) that is involved.

So our rule for 'this' demonstratives goes like this:

> An utterance u of 'This [that] F' refers to b iff
> There is a salience relation *Sal*, that S_u intends to exploit, and
> (a) $Sal(S_u, T_u) = b$
> (b) $F(b)$
> (c) b is relatively near to [far from] S_u at T_u, along the relevant dimension.

(Condition (c) is admittedly a somewhat lame attempt to get at the difference between 'this' and 'that,' but it's the best I can offer.)

When we use simple demonstratives, the kind of object in question is often obvious. If I had merely said, "I like the arches on that," it would have been pretty obvious that I was talking about a building. However, we can use 'this' and 'that' when we don't have much idea what sort of object we are referring to. Dikran and I hear a loud noise on our walk. "What was that?" I ask. "An explosion in a lab? Lightning? A car-crash?" I am asking about the cause of the noise we heard, but it's not clear that I am referring to the noise. "That was a noise" would seem an unhelpful reply on Dikran's part.

Suppose I offer this advice,

> If a philosophical problem keeps you awake at night, *that* problem is a good dissertation topic for you.

What I am saying amounts to:

> if there is a problem *x* such that *x* keeps you awake at night, *x* is a good dissertation topic for you.

We could treat 'that problem' in my remark as an undexical use of 'that problem.' The salience relation *Sal* is (roughly) being something that kept the subject of an episode of thinking awake the night before the time of the thought. You are the subject. The time is supplied by the quantifer. I am saying,

> For any time *t*, if there is a problem *x* such that *Sal*(you, *t*) = *x*, then *x* is a good dissertation topic for you.

9 Direct Reference

Important concepts in the theory of reference emerged in 1960s and 1970s. Saul Kripke (1980) claimed that proper names were *rigid designators* (see Chapter 36, NAMES AND RIGID DESIGNATION). This means that the name picks out the same object in every possible world. If I say, "If Hillary Clinton had been nominated in 2008, then Barack Obama would not have been President in 2009," I have said something true, because in the worlds similar to ours in which Clinton was nominated in 2008, Obama was not President in 2009. If I say, "If Hillary Clinton had been nominated in 2008, then the winner of the 2008 election would not have been President in 2009," I seem to have contradicted myself. Although 'Barack Obama' and 'the winner of the 2008 election' both, in fact, stand for Obama, the latter is not a rigid designator.

David Kaplan introduced a similar concept in *Demonstratives* (Kaplan, 1989), *direct reference*. The idea is that a name *directly* contributes its referent to the *content* of an utterance; rigid designation is the upshot of direct reference. Kaplan maintained that indexicals and demonstratives, like names, are directly referential. The word 'direct' can seem a bit misleading. When I use the indexical 'you' in a conversation with Dikran, I refer to Dikran in virtue of the fact that he is the person I am talking to; the connection between my use of 'you' and Dikran is not direct, but mediated by the fact that I am talkng to him. But Kaplan does not intend "direct" to mean that the *mechanism* of reference is *unmediated*. He means that, whatever complexity may be involved in the mechanism of reference, it is the referent that becomes a constituent of the proposition one expressses, rather than some identifying condition of it. The route from proposition to referent is direct.

Frege's picture contrasts with both Kripke's and Kaplan's. Frege (1892) did not countenance singular propositions; he recognized only qualitative propositions. That is, propositions encode truth-conditions by determing how properties and relations need to be instantiated and co-instantiated for truth; individuals are not among their constituents.[9]

When I say "Bill Clinton was President in 1995," Frege would suppose that I have some identifying condition of Bill Clinton in mind, and the proposition I express is that the person who uniquely satisfies this condition was President in 1995. Even though 'Bill Blythe' and 'Bill Clinton' are both names of Clinton, someone might think that Bill Clinton was President in 1995, but have no idea that Bill Blythe was. So, Frege would reason, it is a mistake to suppose that "Bill Clinton was President in 1995" and "Bill Blythe was President

in 1995" express the same proposition; the sentences have different cognitive significance that should be reflected in the propositions the sentences express.

Kripke, Donnellan (1972), and Kaplan offer convincing arguments that we express singular propositions; that is, such propositions get at what we normally take to be *what is said* by using names and indexicals. This leaves us with a dilemma. If they are right, how do we account for differences in cognitive signficance?

Frege wanted a single notion of content (*Gedanke*) to do two jobs. First, contents are pieces of information passed from person to person, and indeed generation to generation, by the use of language. Second, contents should capture the differences in cognitive significance that come from different ways of thinking of the same thing. But the jobs are different. To get a piece of information, we typically abstract from the different modes of presentation various individuals might have towards the object the information is about. This is what O-conditions get at. Cognitive significance is a matter for S-conditions and E-conditions. Frege's idea that there is a single level of truth-conditions that does both jobs is a mistake (Perry, forthcoming b).

10 A Problem about 'I' and 'Now'

Suppose early on February 16 you see an email from your friend Elwood:

[Preamble]
To: You
Time: February 15, 11 p.m.
From: Elwood Fritchey
[Message]
I am going to bed now.

On the present theory, you know that 'I' refers to the person who wrote the message – call it **m**. Similarly 'now' refers to the time **m** was written. So, just looking at the text, you know the E-conditions of its truth:

m is true iff $\exists x\ \exists t$ such that
$x = R\text{-I}(S_m, T_m)$ &
$t = R\text{-now}(S_m, T_m)$ &
x went to bed at t.

You can then figure out the S-conditions, by consulting the preamble:

m is true iff $\exists x\ \exists t$ such that
$x = R$-I(Elwood, 11 p.m., February 15) &
$t = R$-now(Elwood, 11 p.m., February 15) &
x went to bed at t.

In the case of 'I' and 'now,' the O-conditions follow immediately, at least if we ignore discretionary concerns about 'now':

m is true iff Elwood went to bed at 11 p.m., February 15.

So far, so good. But how about Elwood? Elwood, an English speaker, knows the E-conditions of his email. But how does he know that what he will say, with that email, is that **he** is going to bed at 11 p.m.? He could type the email, and look at the preamble his computer generates, as you did, to figure out which person and time were referred to with 'I' and 'now,' but that doesn't seem necessary.

Elwood might have no idea what day it is, or even what time of day it is, and still be quite sure that when he uses 'now' he is referring to the time at which he is writing and intends to go to bed. It seems he just needs to think of that time as "now." That means he has to think of it as the value of *R*-now for a certain time as argument. But that means, it seems, he must think of *that* time. But how is this to be explained on the present theory? When he thinks "now" it seems that the time he thinks of is the time that is the value of 'now' at the time of the thought. If the only way he can think of this time is as "now" then it seems he identifies the time referred to as

The t such that $t = R$-now(Elwood, now)

But how can this get him any closer to knowing the time to which he is referring? It seems we have a circle, on the verge of becoming a regress.

The solution to this problem lies in what I call "primitive knowledge." Any animal has the ability to pick up information perceptually about the time at which the perception occurs, and use that information in guiding actions taken at the time. A chicken sees a kernel of corn, and pecks at it. In my terminology, the chicken's perceptual state *concerns* the present time, but is not *about* it. The idea is that if we, as chicken-theorists, want to explain the success of the chicken's behavior, *we* would need to *refer* to the time in question. It picked up information about its surroundings *at that time*, and, thanks to the simple architecture of chickens, automatically pecked *at that same time*. The information it picked up perceptually concerned what things were like at a given time, which assured that pecking, *at that same time*, would result in ingesting a kernel of corn. But the chicken does not need to refer to the time at which this all takes place or to think about it. It needs neither indexicals like 'now,' nor the concepts required to distinguish one time from another. Its perceptions automatically provide information about the way things are around it at the time of perception; its perceptual states vary with whether the environment presents an edible kernel or not, but not with the time at which the kernel is presented or not; it is always the time of the perception. The chicken doesn't need to "think" *there is a kernel of corn now*, but simply, *there is a kernel*.

Our lives are a bit more complex than the average chicken. We do have a concept of *now* and the word 'now.' But, we, like the chicken, don't need them and the distinctions that underlie them for simple transactions, in which we act at a given time on information picked up perceptually at that time. We also have primitive knowledge about the present time.

However, unlike the chicken, we do have the requisite concepts and words and can bring them into the picture when needed. We are masters of "the logic of articulation" (Perry, 1986; Korta and Perry, 2011). I am hungry and see a sandwich. I think, *there is a sandwich on my plate*, and eat it. But, if relevant, I infer *there is a sandwich on my plate now*. It is this that makes my concept of *now* a concept of the present time. Similarly, if I know that the bus is coming at 11:15, and see on my watch that it is 11:15, I think *the bus is coming now*, which

in effect makes me think *the bus is coming*, in that it leads me to get ready to board the bus, just as if I had seen it.

So, back to Elwood. Elwood doesn't need to know what time it is to know that he is going to bed. And he doesn't need to know what time it is to know that he is thinking of the time when he thinks *now* and referring to that time if he says "now." He just needs to be aware of that time, in the way one is aware of the present moment, the moment occurring when one is doing the thinking.

Elwood most likely knows that he is Elwood Fritchey. But he might not. Perhaps he is suffering a bout of temporary amnesia. In one sense, he doesn't know to whom he refers with "I," or thinks of when he thinks *I*. But, in another sense, he does. Just as there are special ways of knowing about the present time – look and listen – there are special ways of knowing about yourself. You are the person whose environment you find out about when you look. You are the person whose headaches you feel. And Elwood is the person whose feelings of tiredness he experiences, and the person who forms the intention, as a result of those feelings, to go to bed.

Just as we have primitive knowledge about the present time, we have primitive knowledge about ourselves. The chicken doesn't need to keep track of *which* chicken it is, in order to harness the information it gets from perception to guide its pecking. And Elwood doesn't need to keep track of who is feeling tired, in order to form an intention that will get that very person into bed. So with 'I,' as with 'now,' the threatened regress ends in primitive knowledge. Elwood is the person whose states and environment Elwood learns about in the special ways one knows about oneself. The argument for his use of *R*-I is the person he knows about in this special way.

11 Conclusion

We noted above McTaggart's use of indexicals in his famous argument about time. In fact indexicals, for all their conversational utility and humble origins, seem indispensable in talking about philosophical topics. Try rewriting Descartes's *Meditations* without the first-person, or McTaggart's essay without the words 'past,' 'present,' and 'future.' The reason for this, I suggest, is the connection just considered, between indexicals like 'I' and 'now,' and our primitive but indispensable ways of picking up and using information about the environment we find ourselves in – what is going on around *me, now*. However simple and untechnical 'I' and 'now' may be, compared to words that embody great insights of science or philosophy, they still are words, that express ideas, and connect with complex thoughts words can express. But on the other side, they connect with transactions most indispensable to life: an organism picking up information about itself at a time and putting it to use to its own benefit.

Notes

1 Peirce discusses his distinction in many places. See Atkin (2013).

2 I use double quotes for quoting language and thought, and as scare-quotes; I use single quotes for mentioning expressions, and italics for introducing technical terms.

3 I use *reference* somewhat broadly, rather than confining it to what names and other singular noun phrases stand for.

4 Few such generalizations can withstand the imaginations of philosophers. See, for example, Daniel Dennett's brilliant essay, "Where am I?" (Dennett, 1978).
5 In Kaplan's theory, propositions are sets of world-time pairs, a complication I ignore. An indexical provides a constant intension, that is, a function that picks out the same object in every possible world.
6 Ellen Goodman, *Boston Globe*, April 15, 1996.
7 For an interesting discussion and some useful terminology, see Levy and Olson (1992).
8 On Varol Akman's view the word 'I' has the speaker as a default reference, but in certain circumstances 'I' can refer to others. In my chapter in the previous edition of this *Companion*, I expressed some doubts about his case. However, his distinction between standard and default uses eventually led to the theory of undexical uses developed here. See V. Akman (2002) and Perry (2006).
9 I am being a bit casual here. Frege used the term 'thought' more or less as 'proposition' is used now. He usually eschews talk of "properties" in favor of concepts, which are extensional, and senses, which are not.

References

Akman, V. 2002. "Context and the indexical 'I'." Paper read at the NASSLI '02 Workshop on Cognition: Formal Models and Experimental Results, Center for the Study of Language and Information, Stanford, CA, June 30.

Atkin, A. 2013. "Peirce's theory of signs." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/ (accessed August 27, 2016).

Burch, R. 2014. "Charles Sanders Peirce." In *Stanford Encyclopedia of Philosophy*, edited by Edward E. N. Zalta. Winter 2014. http://plato.stanford.edu/archives/win2014/entries/peirce/.

Castañeda, H.-N. 1999. *The Phenomeno-Logic of the I: Essays on Self-Consciousness*. Edited by J. G. Hart and T. Kapitan. Bloomington and Indianapolis: Indiana University Press.

Dennett, D. C. 1978. "Where am I?" In *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: Bradford Books.

Donnellan, K. 1972. "Proper names and identifying descriptions." In *Semantics of Natural Language*, edited by D. Davidson and G. Harman, pp. 356–379. Dordrecht, Netherlands: Reidel.

Frege, G. 1892. "Über Sinn und Bedeutung." In *Zeitschrift für Philosophie und philosophische Kritik*, 100(1): 25–50. Reprinted as "On sense and reference." In *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edn, edited and translated by P. Geach and M. Black. Oxford: Blackwell, 1980.

Kaplan, D. 1989. "Demonstratives." In *Themes from Kaplan*, edited by J. Almog, J. Perry, and H. Wettstein, pp. 481–614. Oxford: Oxford Univeristy Press.

King, J. 2001. *Complex Demonstratives: A Quantificational Account*. Cambridge: MIT Press.

Korta, K., and J. Perry. 2011. *Critical Pragmatics*. Cambridge: Cambridge University Press.

Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

Levy, D., and K. Olson. 1992. "Types tokens and templates." CSLI Report no. CSLI-92–169. Stanford: Center for the Study of Language and Information Publications.

McTaggart, J. 1908. "The unreality of time." *Mind*, 17(68): 457–474.

Perry, J. 1986. "Thought without representation." *Proceedings of the Aristotelian Society*, suppl. vol. 60: 263–283. Reprinted in *The Problem of the Essential Indexical*, expanded edition, by J. Perry. Stanford: Center for the Study of Language and Information Publications, 2000.

Perry, J. 2006. "Using indexicals." In *The Blackwell Guide to the Philosophy of Language*, 1st edn, edited by M. Devitt and R. Hanley, pp. 314–333. Oxford: Blackwell.

Perry, J. 2011. *Reference and Reflexivity*, 2nd edn. Stanford: Center for the Study of Language and Information Publications.

Perry, J. Forthcoming a. "Indexicals and undexicals." In *Reference and Representation in Thought and Language*, edited by K. Korta and M. Ponte. Oxford: Oxford University Press.

Perry, J. Forthcoming b. "The great detour." In an (as yet unnamed) festschrift for David Kaplan edited by J. Almog and P. Leonardi. Oxford: Oxford University Press.

Predelli, S. 1998. "Utterance, interpretation and the logic of indexicals." *Mind & Language*, 13(3): 400–414.

Reichenbach, H. 1947. "Token-Reflexive Words." In *Elements of Symbolic Logic*. London: Macmillan.

39

Objects and Criteria of Identity

E. J. LOWE

1 Introduction

'Object' and 'criterion of identity' are philosophical terms of art whose application lies at a considerable theoretical remove from the surface phenomena of everyday linguistic usage. This partly explains their highly controversial status, for their point of application lies precisely where the concerns of linguists and philosophers of language merge with those of metaphysicians. The degree of controversy concerning these terms has indeed prompted some skepticism as to their utility (see, for example, Strawson, 1976), but a less pessimistic response would be to try to exercise greater care and discrimination in their use (cf. Lowe, 1989a). Both terms are undeniably slippery, especially 'object.' Our concern will be with the sense of 'object' in which it is interchangeable with 'thing,' but it is important to see that this only coincides with a restricted sense of 'thing.' For we seem to use the word 'thing' in both a narrow and a broad sense, the former associated with the free-standing use of the word and the latter with its use in combination with quantifying adjectives to form unitary quantifier expressions like 'something' and 'everything' (cf. Teichmann, 1992, pp. 15–16, 166–167). The difference is brought out by reflecting on the two non-equivalent sentences 'Every thing is a thing,' which is trivially true, and 'Everything is a thing,' which is metaphysically controversial. (The first sentence means 'Everything which is a thing is a thing,' and is trivial in just the way that 'Every horse is a horse' is trivial; the second sentence, by contrast, is controversial in rather the way that 'Everything is a horse' would be.) As we shall see, some philosophical answers to the question 'What is a thing?' effectively ignore or deny this distinction. My own view is that the distinction is indeed a genuine one, and that it is the narrower sense of 'thing' that is ontologically significant. What is crucial to the status of 'thinghood' in this narrower sense is, I consider, the possession of determinate identity-conditions (see §3 below). This is where the notion of a 'thing' or 'object' ties in with that of a *criterion of identity*, for one guarantee that something possesses determinate

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

identity-conditions is that it falls under a general concept which supplies a definite criterion of identity for its instances. (Such a concept may be classed as a 'sortal.')

As I have already implied, the term 'criterion of identity' is, unfortunately, itself the subject of considerable dispute. One problem is that candidates for this title typically take one or other of two quite different logical forms, whose difference turns on the mode of reference they involve to the objects for whose identity they provide a criterion (see §5). Some objects are such that a canonical mode of reference to them by *functional* expressions of a quite specific kind is available. For instance, to use a famous example of Frege's (Frege, 1953, pp. 74 f.), a particular *direction* may be canonically referred to as the direction *of* a particular line. (Any expression like this, of the form 'the F of x,' may be called a functional expression.) In this particular case the object in question is, of course, not a *physical* but a geometrical one, and this fact may encourage the thought that it is a peculiarity of those objects for which a functional mode of reference is canonical that they are in some sense *abstract* objects, with logico-mathematical objects like directions, shapes, numbers, and sets providing paradigm examples (cf. Dummett, 1981, p. 481). However, as we shall see, the distinction between 'abstract' and 'concrete' objects is itself a highly controversial one, and although it has indeed been argued that this distinction turns ultimately upon differences between the criteria of identity governing objects of these two broad categories (see §10), it certainly does not appear to be simply related to the distinction between those criteria which do and those which do not involve functional modes of reference to the objects they concern. (For one thing, we have indisputably 'abstract' objects like sets, for which a criterion of identity is available which does *not* involve a functional mode of reference to them.) My own view, I should say, is that the distinction I have alluded to between the two types of identity criteria is not, at root, one of fundamental philosophical importance, in the sense of reflecting any basic metaphysical, semantic, or epistemological distinction between the categories of objects to which they apply. This being so, however, one might expect to be able to supplant one or other type of criterion by the other, and I shall indeed try to show how such an expectation may be satisfied in specific cases (see §§7 and 8).

Of course, the very *existence* of abstract objects is itself a matter of considerable philosophical controversy, though it would be inappropriate to engage in it here (but see further Hale, 1987, and Teichmann, 1992, for very contrasting views). However, one should at least be clear as to what is *meant* by 'abstract object' before one debates whether or not anything answers to that description. The putative examples I have so far mentioned – all of them logico-mathematical – are at least provided with clear-cut and unimpeachable criteria of identity; but other putative examples like propositions, facts, and properties do not appear to be so favored. This puts pressure on the idea that propositions and the like possess determinate identity-conditions at all, and correspondingly that they qualify as 'objects' or 'things' (in my narrow sense). That may seem no great loss, until we come to reflect that we can, ostensibly at least, *quantify over* and *refer to* propositions, facts, and properties. However, perhaps we can plausibly represent such 'quantification' and 'reference' as convenient *façons de parler*, capable of being paraphrased away innocuously. I think that is correct, despite the fact that the strategy of paraphrastic elimination is one which must be handled with a good deal of caution, as we shall see (§3). But before we can tackle such issues, we need to examine the role which criteria of identity play in our talk about objects of the least controversial varieties.

2 Sortals and Counting

It is a familiar but nonetheless important philosophical point that an instruction simply to count how many *things* there are in a given room at a certain time is one that cannot be carried out: not because there will always be too many things to count, but because the instruction does not even make determinate sense in the absence of a specification of the *sorts* or *kinds* of things that are to be counted (cf. Geach, 1980, pp. 63 f.; Dummett, 1981, pp. 547 ff.; Wright, 1983, p. 3; Lowe, 1989b, pp. 10 ff.). It makes sense to ask how many *books* there are on a shelf, or how many *girls and boys* there are in a class, because in these cases an appropriate specification is supplied. But what exactly is the nature of such a specification, and what role does it play in conferring determinate sense on such a question? In brief, the point is this. If one is to *count* or *enumerate* items, one must at least be able to *identify* and *differentiate* them, because otherwise some things might be counted more than once. (Just what 'counting' *is* is something that we shall return to later, in §9.) For instance, if I count the books on a certain shelf, I must count each book just once, so that I must be clear as to what differentiates one book from another. A crucial point here is that what differentiates one A from another may not be the same as what differentiates one B from another (where 'A' and 'B' are sortal terms – or, as the linguists appropriately call them, *count nouns* – like 'book' and 'child') – and this is because different sortal concepts supply different *criteria of identity* for the individuals falling under them. A graphic example is provided by an ambiguity in the term 'book' itself, whereby it may either denote a kind of physical object made out of paper, glue, and thread or else a kind of abstract entity possessing certain semantic and syntactic properties. We might call an item of the former kind a 'copy' and an item of the latter kind a 'work.' On a given shelf there might be several copies 'of' the same work, and so the number of books on the shelf in the *former* sense of 'book' would be greater than the number of them there in the *latter* sense.

A further point which emerges from this example, and to which we must return, is that some sortals denote kinds of *concrete* object while others denote kinds of *abstract* object – a distinction of importance, but one whose definition is controversial (see §10). Observe, incidentally, that I spoke above as though items of the abstract kind denoted by the term 'work' might literally occupy a position in space, for instance, a place on a bookshelf; but we shall see later that such talk should perhaps be interpreted in a more roundabout way. (What about *kinds* themselves: are they objects, and if so are they abstract objects? Again, this is something to which we shall return.)

Yet another point emerging from the problem of counting is this: although one must specify what *sorts* of things are to be counted in order to render determinate an instruction to count, it would be wrong to suppose that one can only meaningfully count things of the same kind (cf. Bennett and Alston, 1984; Lowe, 1989b, p. 105). As an earlier example implied, one may count the *boys and girls* in a class, and these are not the same kinds. It is true that boys and girls are both *children*, but that is by the by: one could meaningfully count the *boys and books* in a room, even though there is no single kind (governed by a single criterion of identity) of which both boys and books are sub-kinds. What is crucial is that if one is to count the As and Bs, then (1) A and B must each supply determinate identity conditions for their instances and (2) A and B must be *disjoint* kinds, so that nothing can be an instance of both (for example, one cannot meaningfully be asked to count the *dogs and animals* in a room).

Finally, I should remark that the fact that a general term conveys a criterion of identity for items to which it applies is not a sufficient condition for it to be possible, even in principle, to *count* such items. For *mass* terms like 'gold' and 'water' appear to convey criteria of identity – one can meaningfully ask whether the gold in this room (which might be scattered about it in the form of dust) is *the same* as the gold which formerly composed a certain ornament – even though it makes no sense to ask *how many* gold things, or portions of gold, are currently present in this room, not least because any portion of gold contains other portions of gold as proper parts (see further Simons, 1987, pp. 153 ff.). (By contrast, it does make sense to ask *how much* gold there is in this room.) This shows that a criterion of identity is not exactly the same as a *principle of individuation*, though in the remainder of this chapter we shall chiefly be concerned with count nouns, for which this distinction does not emerge (but see further Woods, 1965). A principle of individuation, we may say, combines a criterion of *identity* with a principle of *unity*: countable items are singled out from others of their kind in a distinctive way that is determined by the sortal concept under which they fall, whereas portions of stuff can only be singled out in *ad hoc* ways, of which there are indefinitely many – as when a portion of gold is singled out as the gold composing a certain ring.

3 What Is an Object?

Of course, not all general terms are sortals, supplying a criterion of identity for items to which they apply: there are also 'adjectival' general terms (Geach) or 'characterizing' universals (Strawson), which supply no such criterion and are, indeed, applicable to things of many different kinds – for example, 'wise' and 'red thing' (see Geach, 1980, pp. 63 f.; Strawson, 1959, p. 168). 'Thing' itself is the most general such term, and is often used interchangeably with the term 'object,' both sometimes being dubbed '*dummy* sortals' (cf. Wiggins, 1980, pp. 63 f.). But what *is* an object, in the most general sense of that term? (I should perhaps stress that what we are seeking here is a satisfactory characterization of what is *meant* by 'object,' not a general criterion for the *existence* of objects of whatever type.) This question is apt to prove confusing. One popular answer, which I shall call the 'Linguistic Answer,' is that anything that can be referred to all – anything that can be made the *reference of a singular term* or be the *value of a variable of quantification* – is 'thing' or 'object' (cf. Frege, 1952a; 1953; Wright, 1983; Quine, 1953a; 1953b). Another possible answer, which I shall call the 'Metaphysical Answer,' is that the term 'object' properly applies to any item which enjoys determinate identity-conditions, and hence any item falling under some sortal concept supplying a criterion of identity for its instances – so that, by this account, a particular *book* (whether a 'copy' or a 'work') or a particular *boy* would qualify as paradigm examples of 'objects.' Now, it may be disputed whether these two answers really *are* different, in the sense of providing different extensions for the term 'object': for it may be contended that anything that can be referred to or quantified over must for that very reason fall under a sortal concept supplying a criterion of identity for its instances. A proponent of the Linguistic Answer endorsing this contention occupies a position which may be epitomized by the two Quinean dicta 'To be is to be the value of a variable' and 'No entity without identity' (see Quine, 1976; 1969; 1990, p. 52). However, the contention in question is certainly open to dispute (cf. Strawson, 1976). For instance, we may apparently refer to the *fact* that such-and-such or the *proposition* that so-and-so, and indeed we may ostensibly quantify over facts and propositions (and likewise over properties, relations, and so forth), but must

we therefore be able to provide *criteria of identity* for such items? It is at the very least highly debatable whether we can, in the light of the interminable philosophical disputes as to what those criteria might be. This is a suspicion which is confirmed by the observation that, although 'fact' and 'proposition' are both grammatically count nouns, there appear to be no principled ways of *enumerating* facts and propositions.

A relevant consideration here may be that apparent reference to and quantification over facts, propositions, and the like seem to be *eliminable by paraphrase*, whence it might be thought that our apparent inability to provide criteria of identity in such cases coincides neatly with the exposure of such 'reference' and 'quantification' as mere *façons de parler* or inflated uses of language. To illustrate these possibilities of paraphrase, instead of saying 'The fact that John was promoted pleased me greatly,' I might less sententiously say 'I was greatly pleased that John was promoted'; and I might paraphrase 'John knows something that I don't' (ostensibly quantifying over propositions) as 'John is somewise more knowledgeable than I am.' (It may be deemed significant that the expression 'somewise' in the latter sentence – admittedly rather an archaism, but none the worse for that – is an *adverb*, in contrast with the *noun* 'something' which figures in the sentence being paraphrased.)

However, there are dangers in putting too much weight upon such possibilities of paraphrase. For one thing, paraphrase is a symmetrical relation, so the fact that reference to or quantification over items of certain kind can apparently be eliminated by paraphrase provides by itself no guide as to *which* of the classes of sentences so related are to be regarded as 'mere' *façons de parler* (cf. Wright, 1983, pp. 25 ff.; but see also Teichmann, 1992, for a defense of the claim that a privileged direction of paraphrase may be discerned). Another point is that it may turn out to be possible to eliminate by paraphrase even reference to and quantification over such paradigm examples of objects as books and children (see, for example, Quine, 1966), but we obviously would not want to say in these cases that such a possibility threatened the status of such items as 'objects.' Certainly this would undermine the suggestion that there is a neat coincidence between cases in which reference to and quantification over items of a certain class are eliminable by paraphrase and cases in which such items cannot be provided with adequate criteria of identity.

The dialectical position we have arrived at now would seem to be as follows. If the Linguistic Answer is combined with an insistence that items referred to or quantified over must be provided with criteria of identity, it looks as though reference to and quantification over facts and propositions must be deemed *ersatz*, since such criteria do not appear to be forthcoming in these cases; however, the possibility of eliminating such reference and quantification by paraphrase provides, it seems, no independent confirmation of the *ersatz* status of such reference and quantification, since such elimination is possible even where criteria of identity *are* available. In the absence of any other independent confirmation, the judgment that such reference and quantification are *ersatz* looks suspiciously like an *ad hoc* maneuver to save the combined view at issue. On the other hand, if the Linguistic Answer is cut free of the demand for criteria of identity, it appears excessively liberal as regards the objects it is prepared to admit to our ontology. The moral which I am inclined to draw is that we should prefer the Metaphysical Answer to the question 'What is an object?' and reject the contention that the Linguistic Answer effectively determines the same extension for the term 'object,' on the grounds that it fails to determine that extension effectively at all. My point would be that the devices of reference and quantification are exploited with immense prodigality in natural language, and resist any principled division into 'genuine'

and 'spurious' (*ersatz*) cases save by appeal to extra-linguistic metaphysical considerations. *Given* the Metaphysical Answer, we are entitled to deny the status of 'objects' to facts and propositions (on the grounds that they lack determinate identity-conditions) and on *this* basis deem 'reference' to and 'quantification' over them mere *façons de parler*, supporting the latter claim by the provision of suitable modes of paraphrase.

4 Frege on Concepts and Objects

It would not do to leave the Linguistic Answer without some further discussion of the views of one of its most esteemed proponents, Gottlob Frege. For Frege, a crucial contrast is to be drawn between objects and *concepts*, the hallmark of the latter being their 'unsaturatedness' (see Frege, 1952a). (The term 'concept' has today a psychological ring which would be quite alien to Frege's intention; in more familiar terminology it may be said to cover both *properties* and *relations*.) In Frege's view, then, the object/concept distinction is a reflection of the linguistic distinction between *subject* and *predicate*. What he has in mind, however, is not the ordinary *grammatical*, but rather the *logical* distinction – the point being that not all grammatical subjects are object-denoting (for example, quantifier phrases, like 'some boy' and 'every book,' are not). So what sort of subject term is object-denoting, on this view? In a word, *names* (*Eigennamen*, in Frege's terminology). However, these must be very broadly construed to include not just ordinary proper names but also definite descriptions (in their 'referential' uses; see Donnellan, 1966), demonstratives, personal pronouns, and so forth. All 'singular terms,' then? Yes, but arguably more besides (even if Frege himself did not think so). For *plural* terms, like 'the books on my shelf' and 'the Joneses,' can function as logical subjects, and surely qualify as object-denoting (see Sharvy, 1980; Boolos, 1984; Lowe, 1991a). Moreover, we should not assume that all object-denoting terms denote *individual* objects, for there are *mass* terms and *kind* terms (like 'gold' and 'tiger,' respectively) which apparently qualify as object-denoting despite not denoting *individuals* (particulars) – rather, they denote *sorts* or *kinds* (of stuff or things; see Lowe, 1989b, pp. 138 ff., 199 ff.; 1991a).

Now sorts or kinds are universals, and therefore presumably *abstract* objects (of which more anon). But what about the adjectival or characterizing universals mentioned earlier: are *they* not likewise objects, at least according to the Fregean view now under examination? This is where we run into Frege's paradox of the concept *horse* (Frege, 1952a): though bearing in mind what I have just said, 'horse' is an ill-chosen example because it is very arguable that 'horse' *does* function as a name and denotes an abstract object, the horse kind; for it can function as a logical subject, as in 'Horses eat grass' and 'Horses are mammals,' which I for one *don't* see (in the way Frege did) as involving quantification over individuals (see Lowe, 1989b, pp. 138 ff.; 1991a). A better example, from my point of view, would be the concept (or, as we might more familiarly say, the property) *wise*. The point then is that '– is wise' functions as *a predicative* expression and so is not object-denoting by Frege's account, because what it expresses is 'unsaturated' (that is, demands 'completion' by an object to form a whole proposition). But if we try to *refer* to what it expresses (by speaking of 'the concept *wise*,' or 'the property of wisdom,' or even just 'what "– is wise" expresses'), then by Frege's own lights we only succeed, it seems, in referring to an *object* which perforce is *not* the concept in question, but a surrogate (see Dummett, 1981, pp. 211 ff.; Palmer, 1988, pp. 36 ff.). Quite what to make of this puzzle is far from clear, though

Frege's own attitude to it (namely, that language prevents us from saying what we want to here, but that we can still somehow get the appropriate message across: Frege, 1952a) certainly does not appear at all satisfactory. I confess I am strongly tempted to see the paradox as an artifact of the Linguistic Answer to the question 'What is an object?' and hence to regard it as a further consideration (though perhaps only a minor one) in favor of the Metaphysical Answer. (According to the Metaphysical Answer, of course, what – if anything – excludes properties like wisdom from the realm of objects is that they lack determinate identity-conditions.)

Rather than pursue this dispute further, however, it may be more profitable to build upon the common ground which clearly exists between an advocate of the Metaphysical Answer, like myself, and most proponents of the Linguistic Answer (namely, those who also subscribe, with Frege and Quine, to the view that reference to and quantification over any class of items presupposes the availability of criteria of identity for those items). This common ground is that to all intents and purposes we can take an object to be any item falling under a sortal concept which supplies a well-defined criterion of identity for its instances. Our next task, then, is to attend to certain difficulties attaching to the very idea of a criterion of identity.

5 Two Forms of Identity Criterion

The notion of a criterion of identity is one which, again, we owe largely to Frege (Frege, 1953, pp. 73 ff.), though we can find antecedents to it in ancient and medieval discussions of the *principium individuationis* (see, for example, Anscombe, 1981; Gracia, 1988) and in Locke's discussion of the idea of identity (Locke, 1975, pp. 328 ff.). Foremost, perhaps, amongst the difficulties attaching to this notion is the question of what *form* such a criterion may or should take. There are two paradigms to be found in the literature, which we may distinguish (using the convenient nomenclature of Timothy Williamson, 1990, pp. 145 ff.) as 'one-level' and 'two-level' identity criteria (see also Lowe, 1989a). Take the example of *sets*. A *one-level* criterion of identity for sets is provided by the Axiom of Extensionality, as follows:

(S1) $\forall x \forall y((\mathrm{Set}(x) \,\&\, \mathrm{Set}(y)) \rightarrow (x = y \leftrightarrow \forall z(z \in x \leftrightarrow z \in y)))$

In words: if x and y are sets, then x is identical with y if and only if x and y have the same members. A *two-level* criterion of identity for sets is provided by Frege's (fatal) Axiom V of the *Grundgesetze* (see Frege, 1952b, pp. 234 ff.; Wright, 1983, p. 155):

(S2) $\forall F \forall G(\{x: Fx\} = \{x: Gx\} \leftrightarrow \forall x(Fx \leftrightarrow Gx))$

In words: the set of Fs is identical with the set of Gs if and only if all and only Fs are Gs. This axiom was the source of notorious difficulty for Frege, because unless a suitable restriction on possible values of 'F' and 'G' is specified, Russell's paradox can be generated from it (see Frege, 1952b). Other well-known Fregean two-level criteria of identity are his criterion of identity for *directions* (Frege, 1953, pp. 74 f.):

(D2) $\forall x \forall y((\mathrm{Line}(x) \,\&\, \mathrm{Line}(y)) \rightarrow (dx = dy \leftrightarrow x // y))$

(the direction of line x is identical with the direction of line y if and only if lines x and y are parallel with one another) and his criterion of identity for *cardinal numbers* (Frege, 1953, pp. 73 f.):

(N2) $\forall F \forall G(Nx\colon Fx = Nx\colon Gx \leftrightarrow \exists R(\{x\colon Fx\}\ 1-1_{R}\ \{x\colon Gx\}))$

(the number of Fs is identical with the number of Gs if and only if the set of Fs is one-to-one correlated with the set of Gs).

The key formal differences between one-level and two-level identity criteria may be described as follows. One-level criteria explicitly quantify over objects of the sort for which they supply a criterion of identity, and state that criterion in terms of a biconditional, one side of which contains a simple expression of identity between such objects and the other side of which expresses an equivalence relation obtaining between those identified objects. By contrast, two-level criteria quantify over items of a *different* kind from that of the objects for which they supply a criterion of identity, and state that criterion in terms of a biconditional, one side of which contains an expression of identity between such objects in which they are referred to by means of *functional* terms relating them to items of the kind quantified over, and the other side of which expresses an equivalence relation obtaining between the items to which the identified objects are thus related.

A difficulty which can beset either form of identity criterion is that of *impredicativity*, which threatens to render such criteria viciously circular. (An impredicative criterion is one which involves "appeal to a totality that includes or depends on" the very objects whose identity is in question: Quine, 1985, p. 166.) It is important to recognize, however, that impredicativity does not *inevitably* give rise to vicious circularity. It doesn't, for instance, in the case of (S1), even if it is advanced in the context of 'pure' set theory of the Zermelo–Fraenkel type, in which all sets save the empty set only have other sets as members (see further Lowe, 1989c). But there certainly *can* be such circularity, as for instance in Donald Davidson's one-level criterion of identity for *events* (see Davidson, 1980; Quine, 1985; Lowe, 1989a; 1989c):

(E1) $\forall x \forall y((\text{Event}(x) \,\&\, \text{Event}(y)) \rightarrow$
$(x = y \leftrightarrow \forall z(\text{Event}(z) \rightarrow ((\text{Cause}(x, z) \leftrightarrow \text{Cause}(y, z)) \,\&\, (\text{Cause}(z, x) \leftrightarrow \text{Cause}(z, y))))))$

In words: if x and y are events, then x is identical with y if and only if x and y cause and are caused by the same events. This is circular inasmuch as what makes for sameness amongst events is precisely what a criterion of identity for events is supposed to convey, and yet a grasp of that is needed in order to understand what is expressed on the right-hand side of the main biconditional in (E1). (This is more obvious when (E1) is expressed in words as above than it is when logical symbolism is employed as in the formula (E1) itself: but there, too, we can see that the repetition of the variable 'z,' understood as taking events as its values, is equivalent to an expression of event-identity.) A similar problem does not beset the criterion of set identity (S1) despite the fact that sets may themselves be set-members, because according to standard set theory, at least, sets belong to a cumulative hierarchy in which (S1) fixes the identity of each set recursively, beginning with sets which contain only non-sets as members, or with just the empty set in the case of 'pure' set theory (see further Lowe, 1989c).

Certain difficulties peculiar to two-level criteria arise from the fact that they utilize *functional* expressions to refer to the objects for which they supply a criterion. One difficulty is that this limits their scope of application quite considerably, at least in the absence of further theorizing. For instance, we need to be able to employ other means of referring to numbers than expressions of the form 'the number of Fs' – not least the numerals '1,' '2,' '3,' and so on. Thus Frege's criterion (N2) doesn't of itself determine the truth-conditions of a statement like '1 + 2 = 3' or 'The number of books on my shelf is eighteen.' Another difficulty is that when we turn away from the sort of mathematical examples which interested Frege, we are often hard put to think of an appropriate two-level way of stating identity criteria. Consider, for instance, the problem of *personal* identity: the trouble is that there is no standard *functional* mode of referring to persons as there is to directions and numbers and sets. Directions are directions *of lines*, and numbers are numbers *of objects satisfying some condition*, as also are sets. But persons aren't at all obviously persons 'of' anything at all in this sense – in short, it isn't obvious what domain of entities ought to be invoked in order that an equivalence relation on *them* may be cited as a criterion of identity for persons (but see Williamson, 1990, pp. 116 ff., for a two-level proposal concerning personal identity).

Even setting aside the foregoing difficulties, which may not seem very serious, it is clear that the two-level approach to identity criteria contains a built-in limitation inasmuch as any such criterion presupposes the identity of items of one kind in providing a criterion of identity for those of another. Thus (D2) presupposes the identity of lines in providing a criterion of identity for directions, and (N2) presupposes the identity of sets in providing a criterion of identity for cardinal numbers. (By saying that (D2) 'presupposes the identity of' lines I mean, of course, that in the absence of a further criterion of identity for *lines* (D2) does not provide a fully informative account of what distinguishes one direction from another.) One-level criteria are not inherently subject to this limitation, which suggests that they will in any case have to be invoked at some stage whenever two-level criteria are themselves invoked. This inevitably provokes a query as to whether two-level criteria are really needed at all, that is, as to whether the work which they do might not be equally well effected by one-level criteria. For unless there are compelling reasons for supposing that two-level criteria provide an indispensable service, considerations of simplicity and parsimony urge us in the direction of regarding one-level criteria as constituting the canonical form. Before we explore this issue, however, one or two preliminary remarks are in order concerning the logical status and role of identity criteria quite generally.

6 The Logical Status and Role of Identity Criteria

The first thing to stress is that criteria of identity are to be thought of, for present purposes, as logico-metaphysical rather than heuristic or epistemic principles – they tell us, in Locke's words, "wherein identity consists" for objects of a given kind (Locke, 1975, p. 335), *not* how we may set about discovering the truth or falsehood of an identity statement concerning such objects; though, obviously, they will not be totally irrelevant to the latter sort of issue (cf. Lowe, 1989b, pp. 15 f.).

Second, identity criteria are not *definitions* – neither of *identity*, nor of *identity restricted to a certain sort or kind* (for identity is univocal), nor even of the *sortal terms* for which they supply criteria (cf. Lowe, 1989b, pp. 22 ff.; Williamson, 1990, pp. 148 ff.). Neither one-level nor two-level identity criteria are apt to provide definitions of the associated

sortals ('direction,' 'number,' and so forth). For two-level criteria, as Frege recognized (Frege, 1953, pp. 77 ff.), do not enable one to replace *all* occurrences of those sortals, only those in which they figure in functional expressions flanking an identity sign on both sides. And one-level criteria involve, as we have seen, reference to and indeed quantification over things of the very sort for which they provide a criterion, and accordingly presuppose some grasp of the associated sortals. (This is made quite explicit in the one-level criteria formulated above – (S1) and (E1) – in which the relevant sortal figures in the antecedent of the formula, instead of a restriction being imposed on the domain of quantification.) So, although it is true that criteria of identity can be construed as conveying semantic information about the sortal terms they relate to (and, certainly, a full grasp of the meaning of those sortal terms requires a grasp of their associated criteria of identity), they do not completely specify the meanings of those terms. This is a fact which, indeed, becomes obvious once it is realized that many *different* sortals are governed by the *same* criterion of identity. ('Cat' and 'dog,' for example, are so governed – for cats and dogs both being kinds of animal, they necessarily both share the criterion of identity governing the sortal 'animal': it would be hard indeed if 'that dog' and 'the animal in that cage' conveyed different identity criteria, given that they may refer to one and the same object.)

Third and finally, I should emphasize that it is not enough for a criterion of identity for As simply to state a logically necessary and sufficient condition for A-identity: it must state such a condition in an informative and, more particularly, a *non-circular* way – by which I mean that a grasp of A-identity must not already be needed in order to understand what is involved in the satisfaction of the condition in question (cf. Lowe, 1989b, pp. 20 f.). As we saw earlier, Davidson's one-level criterion of identity for events, (E1), fell foul of this requirement.

7 One-Level versus Two-Level Identity Criteria

Let us now return to the issue of whether two-level identity criteria are dispensable. One obvious thought is that they may be capable of reformulation in one-level style. (The reverse could not in general be true, in view of our remarks towards the end of §5.) Consider (D2), then, the Fregean criterion of identity for directions, which tells us that the direction of line x is identical with the direction of line y if and only if lines x and y are parallel with one another. Why not reconstrue this in one-level style as the principle that directions are identical just in case any lines of which they are the directions are parallel with one another (cf. Lowe, 1989a)? That is:

(D1) $\forall x \forall y((\text{Direction}(x) \,\&\, \text{Direction}(y)) \rightarrow$
$(x = y \leftrightarrow \forall w \forall z((\text{Line}(w) \,\&\, \text{Line}(z) \,\&\, \text{Of}(x, w) \,\&\, \text{Of}(y, z)) \rightarrow w /\!/ z)))$

It may be objected (cf. Williamson, 1990, pp. 146 f.) that (D1) cannot strictly say the same thing as (D2) because it exploits new terminology in the form of the expression 'Of' (which expresses the relation between a direction and a line of which it is the direction). But, first, exact synonymy is not our target anyway, or else there would be no real advantage in trying to 'reconstrue' two-level criteria in one-level terms; and, second, we might in any case urge that the meaning of 'Of' must be implicitly grasped by anyone who can understand the functional expression 'the direction of x,' which is symbolized in (D2) by 'dx.' Here, however, it may be

further objected that, indeed, 'Of(x, w)' in (D1) is *only* to be understood as a paraphrase for 'dw = x,' so that it is an illusion to suppose that (D1) really dispenses with such functional expressions (cf. Williamson, 1990, pp. 146 f.). But it is not at all clear to me that this suggestion is correct, and we have in any case noted already, in §3, that a possibility of paraphrase does not of itself establish semantic priority (because paraphrase is a symmetrical relation).

Other objections may perhaps be raised against the attempt to reconstrue (D2) as (D1), though I shall not pursue them here (but see further Lowe, 1991c). I must, however, reject Williamson's charge that the Fregean approach of (D2) can, whereas the one-level approach of (D1) cannot, explain why directions and lengths have *different* criteria of identity. According to Williamson, the explanation is that they do so 'because two lines can have the same direction and different lengths, or *vice versa*' (Williamson, 1991, p. 195). But in reality this is no explanation at all, for if it were correct parity of reasoning would require us to say that *heights* and *widths* must have different criteria of identity because two plane figures can have the same height and different widths or *vice versa*, yet heights and widths are both kinds of lengths, being vertical and horizontal lengths, respectively, and so must in fact share the *same* criterion of identity, namely, that of lengths in general. (Observe that this doesn't imply that any height can be *identified* with any width, any more than the fact that cats and dogs share the same criterion of identity implies that any cat can be identified with any dog.) As to the question of what, then, is the correct explanation for the fact that directions and lengths have different criteria of identity, I can only say that the search for an *explanation* of this sort of fact seems to me misplaced from the outset: criteria of identity are built into the very sense of sortal terms, so that to ask why things of the sort which a sortal term denotes are governed by the criterion which it conveys is comparable to asking, absurdly, why the sort of things which it denotes is the sort of things that it is.

The case of directions is not, however, of enough intrinsic importance for too much to hang upon it: Frege himself only introduced it for illustrative purposes. It would be more interesting and potentially fruitful to explore a more fundamental case, such as that of the criterion of identity for cardinal numbers. However, we should bear in mind that what is ultimately at issue is whether two-level identity criteria are dispensable, and to demonstrate that they are it is not necessary to show that they can always be *reconstrued* in one-level terms. Rather, it may suffice to show that we can always supply an adequate one-level criterion *in place of* any two-level criterion; for one criterion of identity is all we need for any given kind of objects, especially if we can also *derive* any correct two-level criterion from an adequate one-level criterion, perhaps with the aid of other necessary truths or definitions. (As we shall see, however, matters may not end quite there, since questions of epistemological and semantic priority may still remain outstanding.)

8 On the Identity of Cardinal Numbers

Consider, then, the case of cardinal numbers. (I should stress that in what follows we shall only be concerned, as Frege himself was, with cardinals no larger than the smallest transfinite cardinal.) What might a one-level criterion to replace Frege's (N2) look like? Here is a possibility:

(N1) $\forall x \forall y((\text{Number}(x) \,\&\, \text{Number}(y)) \rightarrow$
$(x = y \leftrightarrow \forall z(\text{Number}(z) \rightarrow (\text{Precede}(z,x) \leftrightarrow \text{Precede}(z,y)))))$

In words: if x and y are cardinal numbers, then x is identical with y if and only if all and only the cardinal numbers preceding x also precede y (precede, that is, in the series of cardinal numbers <0, 1, 2, 3, …>). Of course, (N1) is 'impredicative' – but only in the harmless way in which (S1) is. No vicious circularity ensues. Criterion (N1) serves to identify 0 unambiguously as the cardinal number which has no predecessors (compare the empty set), and to identify all succeeding cardinal numbers in a recursive fashion (thus 1 is the cardinal number which has as its sole predecessor the cardinal number which has no predecessors, that is, 0, and so on). It is indisputable that (N1) cannot of itself convey the meaning of the sortal term 'cardinal number' to anyone not yet possessed of the concept, and so cannot be taken as providing anything like a definition of this term; but that, as we have seen, should not be regarded as part of the function of a criterion of identity in any case.

An interesting question to raise now is this: Can we recover Frege's principle (N2) from (N1), supplemented with some further necessary truths or definitions? It appears that we can. First we need to define functional expressions of the sort used in (N2), 'Nx: Fx,' – 'the number of Fs.' The obvious thing to say is that the number of Fs is the cardinal number the set of whose predecessors is one-to-one correlated with the set of Fs. More formally, we may adopt the following definition:

(Def N) $\mathrm{Nx}\colon \mathrm{Fx} =_{\mathrm{df}} (\mathrm{iy})(\mathrm{Number}(\mathrm{y}) \,\&\, \exists \mathrm{R}(\{\mathrm{z}\colon \mathrm{Number}(\mathrm{z}) \,\&\, \mathrm{Precede}(\mathrm{z}, \mathrm{y})\}$
$1-1_{\mathrm{R}}\ \{\mathrm{x}\colon \mathrm{Fx}\}))$

In (Def N), I have used 'i' for Russell's definite description operator, so that '(iy)(… y …)' means 'the object y such that … y … ' and is analyzed in Russell's way, according to which (in plain English) 'The object y such that … y … is thus and so' is taken as being equivalent to 'There is one and only one object y such that … y … and y is thus and so.' If in addition to (Def N) we adopt the existence postulate that there is a cardinal number which is the number of Fs, for any condition F (subject to certain necessary restrictions discussed below), that is:

(N*) $\forall \mathrm{F} \exists \mathrm{y}(\mathrm{Number}(\mathrm{y}) \,\&\, \mathrm{Nx}\colon \mathrm{Fx} = \mathrm{y})$

then we are in a position to derive Frege's principle (N2). That is to say, (N1) in conjunction with (Def N) and (N*) entails (N2) (see Appendix 1). Or, in plain English, *given* that cardinal numbers are identical just in case they have the same predecessors, that the number of Fs is the cardinal number the set of whose predecessors is one-to-one correlated with the set of Fs, and that there is a cardinal number which is the number of Fs (and likewise a cardinal number which is the number of Gs), it *follows* that (Frege) the number of Fs is identical with the number of Gs if and only if the set of Fs and the set of Gs are one-to-one correlated.

9 Cardinal Numbers and Counting

But what precisely does the foregoing result serve to show? One's view of that will depend on what semantic and epistemological status one takes Frege's criterion (N2) to have. Is it a principle which has to be grasped by anyone aspiring to a basic knowledge of the cardinal numbers and so of elementary arithmetic? It is not clear to me that it is (but see Wright, 1983, pp. 117 ff., where an opposing view is expressed). Consider this: when children begin to learn about number they do so by learning to *count*. But what is 'counting'? It is a process

of establishing a one-to-one correlation between a set of objects (for instance, the books on a certain shelf) and the set of predecessors of a certain cardinal number: a task which is accomplished by singling out each object just once (often by pointing to it) and uttering a numeral in sequence until every object has been accounted for. In practice, of course, we don't say 'zero' but rather 'one' as we point to the first object, but that is purely a matter of convention: the upshot is still that when we have finished the counting process we 'reach' a number which is the number of the objects being counted, in the sense just defined – that is, a number the set of whose predecessors is one-to-one correlated with the set of objects in question. It is arbitrary whether by 'reaching 3' we mean uttering the sequence of numerals '0,' '1,' '2' or, as is conventional, uttering the sequence of numerals '1,' '2,' '3.' Now, counting provides us with a means whereby to establish the *equinumerosity* of two sets of objects – for example, the books on a shelf and the children in a class – relying on the fact that one-to-one correlation is transitive. Such equinumerosity *can* sometimes be established directly (for instance, by giving each child one and only one book), but often this is impractical. It seems to me that the realization that one-to-one-correlated sets of objects are *equinumerous* is a more sophisticated achievement than the simple ability to *count* sets of objects, and consequently that we should not expect a grasp of Frege's criterion of identity for cardinal numbers to lie at the heart of our basic understanding of number. Indeed, it is a possible objection to Frege's approach that it gives no immediate insight into the relationship between cardinal numbers and the process of counting which is central to a child's induction into a grasp of the numbers (for an extended discussion of this and related matters, not always consonant with the views expressed here, see Dummett, 1991, pp. 143 ff.).

This discussion of counting takes us back to some of the issues of §2. We remarked there that we can only meaningfully be asked to *count* objects when supplied with appropriate sortal specifications. We can now see more clearly why this is so. Counting a set of objects is a process of establishing a one-to-one correlation between those objects and the set of predecessors of a certain cardinal number, which is then designated as the number of that set of objects. But this process demands that each object is identifiable and differentiable from the others, and supplying a criterion of identity for each such object (which is what a sortal specification will convey) normally enables this demand to be met. However, this should not be taken to preclude us from saying that *there are* objects that are uncountable even in principle: for example, the portions of gold, or the red things, currently present in this room. Incidentally, I remarked earlier that a restriction would have to be placed upon the postulate that for any condition F there is a cardinal number which is the number of objects satisfying that condition ((N*) of §8). One reason why this is so is now clear: unless 'F' supplies a concept conveying a criterion of identity for each object falling under it, we cannot meaningfully assign those objects a number. Thus, where 'F' means 'book on this shelf,' there is no difficulty in supposing that there is a number which is the number of Fs: but not so where 'F' means 'red thing currently present in this room.' Observe, though, that even if 'F' *does* supply a concept conveying a criterion of identity for objects falling under it, this does not guarantee that there is such a thing as the number of Fs. For instance, 'set' supplies such a criterion in the form of (S1), and yet we know that there are 'too many' sets for there to be a number of them (though there may, of course, be a number of sets *meeting some further specified condition*, such as the number of 13-membered sets of cards that can be dealt from a 52-card pack: see, for example, Moore, 1990, pp. 147 ff.). Again, there is a criterion of identity governing portions of gold, and yet, as we saw in §2, no number can meaningfully be assigned to the portions of gold currently present in this room

(because mass terms like 'gold' fail to supply a principle of *unity* for their instances). So, stating an appropriate restriction on 'F' in (N*) is no simple matter. How best to handle this problem I shall discuss no further here, beyond saying that one obvious strategy which will serve the purposes to which we put (N*) earlier is to replace (N*) by:

(N**) $\forall F(\exists G \exists R(\{x: Fx\}\ 1 - 1_R\ \{x: Gx\}) \rightarrow \exists y(\text{Number}(y) \,\&\, Nx: Fx = y))$

In words: if the Fs are one-to-one correlated with the members of some set, then there is a cardinal number which is the number of Fs.

10 Abstract and Concrete Objects

One important issue which I have postponed until now is that of the distinction between 'abstract' and 'concrete' objects. I assume that numbers, sets, and directions are uncontroversially abstract, while books and children are indisputably concrete. Of course, it may be asked how I *know* that numbers are abstract, when nothing I have said about them so far determines what they are. Indeed, it has been argued that numbers *could not* be 'objects' at all (see Benacerraf, 1983; but see also Wright, 1983, pp. 117 ff., for criticism). My own view is that the natural numbers, at least, are *sorts* or *kinds* (of sets) and so *a fortiori* abstract (see Lowe, 1993). However, even if this is not accepted, perhaps we know enough about numbers to know that they would have to be abstract whatever they are – perhaps because there are too many of them for them to be concrete.

An obvious suggestion is that concrete objects are, while abstract objects are not, denizens of space-time (or, which perhaps amounts to the same thing, are/are not subject to causality: see, for example, Grossmann, 1992, p. 7). This has been queried, for instance by Bob Hale (1987, p. 49), on the grounds that objects such as *languages* are plausibly abstract and yet come into existence and undergo change and so presumably exist in time. (It won't do to classify them as abstract on the grounds that they *only* exist in time and not also in space – even if it were altogether plausible to say this of them – for we should want to classify Cartesian egos as 'concrete' despite ascribing only temporal, not spatial, existence to them.) Hale proposes instead, developing a suggestion of Harold Noonan's (see Noonan, 1976; 1978), that abstract objects can be distinguished by reference to certain features of the criteria of identity which govern them. Specifically, he proposes (Hale, 1987, p. 61):

(A4) F is an abstract sortal iff, for any R that grounds F, either
- R cannot hold between spatially located items at all or
- (R can hold between things which are spatially, but not temporally, separated

where R is an equivalence relation and R *grounds* F iff, for any statement of identity linking F-denoting terms, there is some statement to the effect that R holds among certain things, the truth of which is (logically) necessary and sufficient for the truth of that statement of F-identity (Hale, 1987, p. 59).

As an example of a grounding relation, Hale cites the relation of parallelism between lines, which qualifies as such 'in virtue of the fact that lines have identical directions iff they are parallel' (Hale, 1987, p. 59). From this it appears that Hale is thinking primarily in terms of two-level ('Fregean') rather than one-level identity criteria; though he acknowledges that

at least some sortals must be governed by one-level criteria (p. 57), and it is clear, indeed, that he intends (A4) to prescind from the distinction between one-level and two-level criteria.

Limitations of space prevent me from discussing the interesting reasoning behind Hale's ingenious proposal, but it appears in any case to be fatally flawed. This is most easily seen if one considers what it implies about *concrete* sortals (assuming that a sortal is 'concrete' if and only if it is not 'abstract'). Negating the right-hand side of (A4), we see that by Hale's account a sortal F qualifies as concrete iff there is some R that grounds F such that (i) R can hold between spatially located items and (ii) R cannot hold between things which are spatially, but not temporally, separated. Now consider the relation 'x and y coincide in their boundaries.' This is clearly a relation which serves to 'ground' the abstract sortal 'part of a geometrical figure,' for it is evident that if x and y are parts of a geometrical figure (for example, semicircular parts of a circle), then they are, of logical necessity, identical parts if and only if they coincide in their boundaries. However, this is a relation which *can* also hold between spatially located items (for instance, Switzerland coincides in its boundaries with itself), but cannot hold between things which are spatially separated (and so *a fortiori* cannot hold between things which are spatially, but not temporally, separated). By Hale's account, therefore, the sortal 'part of a geometrical figure' is wrongly classified as concrete.

However, rather than attempt to refurbish Hale's proposal, let us look again at the previous suggestion that abstract objects are those that are not denizens of space-time. The supposed difficulty was that objects like languages are plausibly abstract and yet also plausibly come into existence and undergo change. But perhaps we need to make a distinction, which can best be brought out by analogy with a related case: that of biological species. These too are said to come into existence and undergo change – indeed, that they do so is crucial to the theory of evolution. How then can species names denote universals, which are abstract entities and so, on the present proposal, timeless? The solution is to distinguish between biological *species*, which are *concrete individuals* consisting at any time of the mereological sum of their currently existing members (particular tigers or particular oaks), and biological *sorts* or *kinds*, which are universals instantiated by the members of those species (see Lowe, 1991a; and cf. Hull, 1976). Thus we can say that the horse *species* at one time did not exist and has evolved over millennia as its individual members have gradually taken on different morphological features, but that the *kind* horse which all these past and present individual horses instantiate never 'came into' existence and has not itself undergone change. In like manner, we may say that 'English,' construed as denoting a kind of language, does not refer to an ephemeral and changeable entity, but that what have come and gone and been subject to change are the concrete processes of linguistic communication which, over the centuries of English history, have all qualified as manifestations of English. On this view, inasmuch as 'English' denotes something abstract it denotes a *kind* (a universal), not an individual. To the extent that we happily identify various *sub-kinds* of English – such as American English and Old English – this view seems reasonable, since only kinds (not individuals) can have sub-kinds.

11 The Paradoxes of Identity over Time

This is a convenient place to address a final issue, which concerns the problem of *identity over time* and the paradoxes to which identity criteria often appear to give rise when time is brought into the picture. (There are also analogous *modal* paradoxes, which, however,

I shall not discuss here; but see Lowe, 1986; and Williamson, 1990, pp. 126 ff., as well as Chapter 40, RELATIVE IDENTITY.) The paradoxes arise because the identity criteria that we are intuitively led to adopt for various kinds of objects which persist through time permit these objects to change in certain respects while remaining numerically the same objects, and yet a series of small and acceptable changes can add up to a large change which we may intuitively feel to be incompatible with the retention of numerical identity for the object concerned. (Such paradoxes are, then, ostensibly a variety of sorites paradox: see Chapter 28, SORITES.)

In short: identity over time must be a *transitive* relation, and yet our intuitive identity criteria for objects persisting through time seem to rely on relations which are not strictly transitive.

For instance: we want to allow that a *ship* can persist identically through small changes in its component parts or in its overall design or structure, but a great many such successive changes may transform it into an object made of completely different materials put together in a completely different way; so that what we eventually have is no longer a ship at all, and so *a fortiori* not the *same* ship as the one we started with. Similar points have been made about *languages* (ignoring for the moment the bearing of my earlier denial that these may literally undergo change when conceived of as abstract entities). For example (see Williamson, 1990, p. 137), the language now spoken in Rome has, we may suppose, developed by small step-by-step changes from the language which was spoken in ancient Rome, such that no one of those changes amounted to the extinction of one language and the birth of a new one; and yet modern Italian is not numerically the same language, surely, as ancient Latin.

To cope with these problems we might attempt to refurbish what we take to be the intuitive identity criteria for artifacts like ships and languages, substituting strictly transitive relations for the non-transitive ones supposedly causing the trouble (cf. Williamson, 1990, pp. 139 ff.). But before taking such drastic action we should explore the possibility that the problems are spurious ones, arising from a confusion between the identity criteria for individuals falling under given sortal concepts and the conditions for the correct application of those sortals to individuals. We need, I suggest, to allow for the possibility of *metamorphosis* (see also Lowe, 1989b, pp. 103 f.), that is, a process whereby one and the same individual object can persist through a transformation from being an object of one sort A to being an object of another sort B, such that no object can *simultaneously* be both an A and a B. A logical restriction on such change is that A and B should supply the same criterion of identity for individuals instantiating them. But we have already noticed that very different sortal concepts can indeed convey the same identity criteria – for instance, the concepts *cat* and *dog* – and, indeed, that all sortals falling under the same higher-level sortals (as *cat* and *dog* both fall under *animal*) must, on pain of contradiction, supply the same identity criteria for individuals instantiating them. There can thus be no *logical* objection to the possibility of an individual animal surviving a change from being a cat to being a dog, even if such a transformation is *physically* impossible for biological reasons. In the case of artifacts like ships and languages such physical restraints are absent, and hence 'metamorphosis' may be expected to be a more common phenomenon amongst them. Thus, we can consistently react to the Italian/Latin example discussed earlier by saying that the same *individual language* has persisted identically in Rome from ancient to modern times, but that in the course of history it has changed from being an instance of the language-type *Latin* to being an instance of the language-type *Italian*, where these language-types are defined by certain

important lexical and syntactic features. (It should be observed that this reaction is consistent with my earlier proposal that to the extent that language-names like 'Latin' and 'Italian' denote abstract entities, they denote kinds or types rather than individuals; furthermore, it may be conceded that the boundary between Latin and Italian is not a sharp one, and even that some sub-kinds of Latin equally qualify as sub-kinds of Italian.) Similarly, one and the same individual artifact might change from being a ship to being a hotel, provided both sortals convey the same criterion of identity.

If this solution is correct, the lesson would be that it is an error to suppose that the criterion of identity for, say, artifacts of a given sort necessarily embodies within it a condition to the effect that such an individual can only persist *as an individual of that sort*. We need to distinguish between the diachronic *identity* conditions of individuals and the conditions for their persistence as individuals *of a given sort* (what we might call 'sortal persistence conditions'; cf. Lowe, 1991b, pp. 93 f.).

Once this distinction is drawn, I surmise, many of the supposed temporal paradoxes of identity will dissolve, since they present no challenge to the transitivity of identity, and only serve to demonstrate that 'metamorphosis' is possible and, indeed, quite common. (There do exist puzzle cases, like that of the ship of Theseus, which genuinely concern identity and cannot be handled in the way just proposed, but I believe that most such puzzles are independently soluble in a quite straightforward fashion: see Lowe, 1983.) Of course, it may be said that we were already familiar with the possibility of metamorphosis from the case of transformations like that of a caterpillar into a butterfly or that of a tadpole into an adult frog: but in fact such transformations are not true cases of metamorphosis as I presently understand that term, because count nouns like 'caterpillar' and 'tadpole' – like also 'boy' and 'sapling' – are what Wiggins has called phased sortals, describing an individual as it is during one period of its natural development (see Wiggins, 1980, p. 24). True metamorphosis, such as that of a cat into a dog or that of a human being into a frog, would not be a natural process; nor can 'cat' and 'human being' properly be called phased sortals.

Appendix: Informal Proof of (N2)

We want to show that (N2) follows from the conjunction of (N1), (Def N), and (N*) (see §8). Suppose, then, that

(1) $Nx: Fx = Nx: Gx$

that is, the number of Fs is identical with the number of Gs. Then, by (Def N), this implies that there is a number y, the set of whose predecessors is one-to-one correlated with the set of Fs, and a number w, the set of whose predecessors is one-to-one correlated with the set of Gs, and $y = w$. If $y = w$, then by (N1) y and w have exactly the same predecessors, and since we are given that the set of these predecessors is one-to-one correlated both with the set of Fs and with the set of Gs, it follows by the transitivity and symmetry of one-to-one correlation that the set of Fs is one-to-one correlated with the set of Gs, that is:

(2) $\exists R(\{x: Fx\}\ 1-1_R\ \{x: Gx\})$

So (2) follows from (1), and hence (N2) holds in the left-to-right direction. Next assume for the converse that (2) is true. Now, by (N*) we have that there is a number y, which is the

number of Fs, and a number w, which is the number of Gs. That is to say, by (Def N), we have that there is a number y, the set of whose predecessors is one-to-one correlated with the set of Fs, and also a number w, the set of whose predecessors is one-to-one correlated with the set of Gs. But by (2) we have that the set of Fs is one-to-one correlated with the set of Gs, whence it follows by the transitivity and symmetry of one-to-one correlation that the set of y's predecessors is one-to-one correlated with the set of w's predecessors. From this it follows that y and w have exactly the same predecessors, and consequently by (N1) that y and w are the same number. But y and w are, respectively, the number of Fs and the number of Gs, which are therefore also identical. So that (2) follows and consequently (N2) holds in the right-to-left direction. *QED*. (Note that the proof will equally go through with (N**) of §9 replacing (N*). It is crucial to the proof, incidentally, that – as stated at the beginning of §8 – we are concerned with cardinals no larger than the smallest transfinite cardinal (for background information see Moore, 1990, pp. 147 ff.))

References

Anscombe, G. E. M. 1981 (1953). "The principle of individuation." In *From Parmenides to Wittgenstein: Collected Philosophical Papers*, vol. 1. Oxford: Blackwell.

Benacerraf, P. 1983 (1965). "What numbers could not be." In *Philosophy of Mathematics: Selected Readings*, 2nd edn, edited by P. Benacerraf and H. Putnam, pp. 272–294. Cambridge: Cambridge University Press.

Bennett, J., and W. Alston. 1984. "Identity and cardinality: Geach and Frege." *Philosophical Review*, 93(4): 553–567.

Boolos, G. 1984. "To be is to be a value of a variable (or to be some values of some variables)." *Journal of Philosophy*, 81(8): 430–449.

Davidson, D. 1980 (1969). "The individuation of events." In *Essays on Actions and Events*. Oxford: Clarendon Press.

Donnellan, K. 1966. "Reference and definite descriptions." *Philosophical Review*, 75(3): 281–304.

Dummett, M. 1981. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.

Dummett, M. 1991. *Frege: Philosophy of Mathematics*. London: Duckworth.

Frege, G. 1952a (1892). "On concept and object." In *Philosophical Writings of Gottlob Frege*, translated by P. T. Geach and M. Black. Oxford: Blackwell.

Frege, G. 1952b. "Frege on Russell's paradox [*Grundgesetze der Arithmetik*, vol. II, appendix]." In *Philosophical Writings of Gottlob Frege*, translated by P. T. Geach and M. Black. Oxford: Blackwell.

Frege, G. 1953 (1884). *The Foundations of Arithmetic*, translated by J. L. Austin. Oxford: Blackwell.

Geach, P. T. 1980. *Reference and Generality*, 3rd edn. Ithaca: Cornell University Press.

Gracia, J. J. E. 1988. *Introduction to the Problem of Individuation in the Early Middle Ages*, 2nd edn. Munich: Philosophia.

Grossmann, R. 1992. *The Existence of the World: An Introduction to Ontology*. London: Routledge.

Hale, B. 1987. *Abstract Objects*. Oxford: Blackwell.

Hale, B., and C. Wright. 2001. "To bury Caesar." In *The Reason's Proper Study: Essays Towards a Neo-Fregean Philosophy of Mathematics*, edited by B. Hale and C. Wright, pp. 335–396. Oxford: Oxford University Press.

Hull, D. L. 1976. "Are species really individuals?" *Systematic Zoology*, 25(2): 174–191.

Locke, J. 1975 (1690). *An Essay Concerning Human Understanding*, edited by P. H. Nidditch. Oxford: Clarendon Press.

Lowe, E. J. 1983. "On the identity of artifacts." *Journal of Philosophy*, 80(4): 220–232.

Lowe, E. J. 1986. "On a supposed temporal/modal parallel." *Analysis*, 46(4): 195–197.
Lowe, E. J. 1989a. "What is a criterion of identity?" *Philosophical Quarterly*, 39(154): 1–21.
Lowe, E. J. 1989b. *Kinds of Being: A Study of Individuation, Identity and the Logic of Sortal Terms*. Oxford: Blackwell.
Lowe, E. J. 1989c. "Impredicative identity criteria and Davidson's criterion of event identity." *Analysis*, 49(4): 178–181.
Lowe, E. J. 1991a. "Noun phrases, quantifiers, and generic names." *Philosophical Quarterly*, 41(164): 287–300.
Lowe, E. J. 1991b. "Real selves: persons as a substantial kind." In *Human Beings*, edited by D. Cockburn, pp. 87–107. Cambridge: Cambridge University Press.
Lowe, E. J. 1991c. "One-level versus two-level identity criteria." *Analysis*, 51(4): 192–194.
Lowe, E. J. 1993. "Are the natural numbers individuals or sorts?" *Analysis*, 53(3): 142–146.
Moore, A. W. 1990. *The Infinite*. London: Routledge.
Noonan, H. W. 1976. "Dummett on abstract objects." *Analysis*, 36(2): 49–54.
Noonan, H. W. 1978. "Count nouns and mass nouns." *Analysis*, 38(4): 167–172.
Palmer, A. 1988. *Concept and Object*. London: Routledge.
Quine, W. V. O. 1953a (1948). "On what there is." In *From a Logical Point of View*. New York: Harper and Row.
Quine, W. V. O. 1953b. "Logic and the reification of universals." In *From a Logical Point of View*. New York: Harper and Row.
Quine, W. V. O. 1966. "Variables explained away." In *Selected Logic Papers*. New York: Random House.
Quine, W. V. O. 1969 (1958). "Speaking of objects." In *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Quine, W. V. O. 1976 (1939). "A logistical approach to the ontological problem." In *The Ways of Paradox and Other Essays*, rev. edn. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1985. "Events and reification." In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, edited by E. Lepore and B. McLaughlin, pp. 162–171. Oxford: Blackwell.
Quine, W. V. O. 1990. *Pursuit of Truth*. Cambridge, MA: Harvard University Press.
Sharvy, R. 1980. "A more general theory of definite descriptions." *Philosophical Review*, 89(4): 607–624.
Simons, P. 1987. *Parts: A Study in Ontology*. Oxford: Clarendon Press.
Strawson, P. F. 1959. *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.
Strawson, P. F. 1976. "Entity and identity." In *Contemporary British Philosophy, Fourth Series*, edited by H. D. Lewis, pp. 193–219. London: George Allen and Unwin.
Teichmann, R. 1992. *Abstract Entities*. Basingstoke: Macmillan.
Wiggins, D. 1980. *Sameness and Substance*. Oxford: Blackwell.
Williamson, T. 1990. *Identity and Discrimination*. Oxford: Blackwell.
Williamson, T. 1991. "Fregean directions." *Analysis*, 51(4): 194–195.
Woods, M. J. 1965. "Identity and individuation." In *Analytical Philosophy, Second Series*, edited by R. J. Butler, pp. 120–130. Oxford: Blackwell.
Wright, C. 1983. *Frege's Conception of Numbers and Objects*. Aberdeen: Aberdeen University Press.

Postscript

HAROLD NOONAN

This postscript takes its starting point from the common ground established in the main text (§4): that we can take an object to be any item falling under a sortal concept which supplies a well-defined criterion of identity for its instances. So our topic is the notion of a criterion of identity. David Lewis famously denied that there are problems about identity: "We do state … genuine problems in terms of identity. But we needn't state them

so. Therefore, they are not problems about identity" (Lewis, 1986, p. 193). In this postscript we provide an account, conforming to Lewis's dictum, of what one is providing when one gives a criterion of identity for objects of sort S.

There are, as noted in the main text, two forms of identity criteria in which an equivalence relation between objects (or entities of a higher level, concepts, for example) is specified as the criteria relation for the sort of object for which the criterion is being stated: one-level and two-level. A one-level criterion of identity for objects of sort S takes the form:

> If x is an S and y is an S then x = y iff Rxy
> for example, 'If x and y are sets then x is identical with y iff x and y have the same members.'

A two-level criterion for Ss takes the form (restricting ourselves to examples in which the criterial relation is one holding between objects):

> If x is an S* and y is an S* then dx = dy iff Rxy,
> For example, 'If x and y are lines then the direction of x is identical with the direction of y iff x and y are parallel.'

The key formal differences are noted in §5.

A two-level 'criterion of identity' is thus in the first place an implicit definition of a functor 'd' (e.g., 'the direction of') in terms of which a sortal predicate 'is an S' (e.g., 'is a direction') can be defined ('is a direction' = 'is the direction of some line'). Consistently with the two-level criterion of identity stated several distinct functions may be the reference of the functor 'd.' Hence, as stated in the text, two-level criteria are neither definitions of identity, nor of identity restricted to a certain sort (for identity is universal), nor even of the sortal terms for which they provide criteria. They merely constrain, but not to uniqueness, the possible referents that the functor 'd' represents and thus give a merely necessary condition for falling under the sortal predicate 'is an S' (when 'x is an S' is explained as equivalent to 'for some y, x is identical with dy').

Moreover, as noted in the text, and by Hale and Wright (2001), two-level criteria may be replaced by equivalent one-level criteria.

A one-level criterion of identity for things of sort S of the form above is equivalent to the conjunction of:

> If x is an S then Rxx

and

> If x is an S then if y is an S and Rxy then x = y.

Each of these gives a merely necessary condition for being an S (the second may also be viewed as giving, non-circularly, a necessary condition for a sortal concept to *be* the concept of an S: that there are not two R-related objects falling under that concept).The second of these says something about Ss which is not true of everything only if 'Rxy' does not entail 'x = y.'

Together these are equivalent to the proposition that every S is *the* S R-related to it. (Note that it may be that every S is the S R-related to that S, but it need not be the case, even though

every S is an S*, that every S* is the S* R-related to that S*, because some S*s are not R-related to themselves at all, or, even if Ss, are R-related to other S*s (see Cartwright, 1967; Noonan, 2009), so conclusions about non-membership of sorts (no Ss are S*s, not every S is an S*, e.g., no sets are properties, not every set is a property) cannot be safely drawn from premises about the distinctness of one-level criteria of identity.) The one-level criterion of identity for things of sort S thus again merely specifies a necessary condition for being an object of sort S.

A third possibility, not discussed in the main text, remains. This is that there are criteria of identity which are neither one-level nor two-level, but as it were, *zero*-level, because grasp of them does not consist in associating a criterial *relation* between objects (or any higher-level entities) with a sortal term at all. This is Dummett's proposal (1981; 1991). To understand a basic sortal term, like 'cat,' one must understand a statement of identification of the form 'this is the same cat as that' (uttered, for example, pointing first towards an ear and then a tail). But a statement of identification does not of itself involve any reference to objects since it is merely a crude relational predication (like 'this is darker than that') in which 'is the same cat as' does not stand for any relation at all. Reference to, and quantification over, objects takes place only at a higher level of language, and it is only at the higher level that one-level and two-level criteria of identity can be stated.

In addition to questions about criteria of identity *simpliciter* there are, for objects of some sorts, questions about criteria of identity over time.

It is a philosophical platitude, going back to Locke, that conditions of identity over time are sort-dependent. That is, they vary with the sort of object in question: "it being one thing to be the same person, another the same man, another the same substance, if *person*, *man* and *substance* stand for three different ideas" (Locke, 1975, essay II, xxvii.7). But as we will see, the specification of the conditions of identity over time for objects of sort S is just the specification of certain necessary conditions on being an S, those which constitute certain constraints on the history of a thing of sort S.

Recall the familiar philosophical puzzle of the statue and (sometime) coincident piece of clay. The usual argument for non-identity is that each can survive changes the other cannot. Our concept of a statue plausibly implies that *no statue can survive radical reshaping*. Our concept of a piece of clay plausibly implies that *any piece of clay must survive radical reshaping if all its matter is preserved in one coherent whole*. These propositions specify persistence conditions for statues and pieces of clay and, as they illustrate, these persistence conditions are of two types.

The proposition that no statue can undergo radical reshaping can be expressed as:

> If x is a statue then if the matter that constitutes x at t is radically reshaped at t, then x ceases to exist

– this specifies a 'passing away' condition for statues (this terminology comes from Penelope Mackie, personal communication).

The proposition that any piece of clay must survive radical reshaping in which all its matter is preserved in one coherent whole can be expressed as:

> If x is a piece of clay then if the matter that constitutes x at t is radically reshaped at t but preserved in one coherent mass, x survives

– this specifies a 'preservation' condition for pieces of clay.

Sortal concepts for persisting things are governed by such conditions because they constrain the histories of the things they apply to, and such constraints can always be expressed in the form:

If x is an S then if x exists at t and t* then Rxtt*

or in the form:

If x is an S then if Rxtt* and x exists at t then x exists at t*.

The 'passing-away' condition for statues is entailed by a principle of the first form (stating that a statue cannot have radically different shapes at different times) and the 'preservation condition' for pieces of clay is entailed by a principle of the second form (stating that if the matter composing a piece of clay at one time is in one coherent mass at that time and at another, whatever shape it is in the piece of clay exists at both times). We may call the first principle a 'passing away' principle and the second principle a 'preservation' principle.

Principles of these forms state necessary conditions for being a thing of sort S. A principle of the first form rules out certain changes in the history of an S. A principle of the second form rules out something's being an S if its history is not extensive enough; it tells us that if something is an S and there occur appropriately related events, one occurring at a time it exists, the second must also. Thus a principle of the second form imposes a certain 'maximality' condition on the concept of an S. What distinguishes sortal from non-sortal concepts under which persisting things fall (even from ones that apply to a thing at any time it exists, like *permanent bachelor*) is that they are governed by such principles.

So questions about criteria of identity over time can be rephrased as questions about necessary conditions of membership in a sort. One question we can ask, without mention of identity, is:

(Q1) What conditions *R* satisfy the following schema: (P1) if x is an S then if x exists at t and t*, Rxtt*?

Another is:

(Q2) What conditions R satisfy the following schema: (P2) if x is an S then if Rxtt* and x exists at t then x exists at t*?

To explain the criterion of identity over time for things of sort S is just to answer these two questions.

Thus all intelligible questions about the criterion of identity of things of sort S are equivalent to questions about the necessary conditions for being an object of sort S, in conformity with Lewis's dictum.

References

Cartwright, R. L. 1967. "Classes and attributes." *Noûs*, 1(3): 231–241.
Dummett, M. 1981. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.

Dummett, M. 1991. *Frege: Philosophy of Mathematics*. London: Duckworth.

Hale, B., and C. Wright. 2001. "To bury Caesar." In *The Reason's Proper Study: Essays Towards a Neo-Fregean Philosophy of Mathematics*, edited by B. Hale and C. Wright, pp. 335–396. Oxford: Oxford University Press.

Lewis, D. K. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Locke, J. 1975 (1690). *An Essay Concerning Human Understanding*, edited by P. H. Nidditch. Oxford: Clarendon Press.

Noonan, H. W. 2009. "What is a criterion of identity?" *Analysis*, 69(2): 274–277.

40

Relative Identity

HAROLD NOONAN

Introduction

A piece of bronze is shaped into a statue of Napoleon and then some time later melted down and shaped into a statue of Winston Churchill. Thus the same *piece of bronze* is, at different times, different *statues*. A ship undergoes a process of repair and replacement of parts so that eventually not a plank of the original remains. Thus the same *ship* is at different times two completely different *collections of planks*. Dr Jekyll drinks his potion and transforms himself into Mr Hyde. Thus the same *man*, at different times, is two different *persons* or *personalities*. In the zoo there are several individuals of the same species: Tiger Tim is the same *species of animal* as Tiger Tom, but a different *member* of the species. In the Trinity, the Father, the Son, and the Holy Ghost are the same *God*, but three different *Persons*.

These examples suggest that one and the same A may be different Bs, and hence that there is some sort of incompleteness or indefiniteness in the unqualified statement that x and y are the same, which needs to be eliminated by answering the question 'the same *what?*'

One way of making these vague thoughts more precise is by appeal to Geach's idea that *identity is relative* (see Geach, 1962; 1967 and subsequent references). Major contributions on the opposing side were those of Wiggins (1967; 1980) and Dummett (1973; 1981; 1991); views similar to Geach's were put forward by Quine (1963; 1973), Chisholm (1969; 1970; 1976), and Lewis (1976). In what follows I evaluate Geach's main claims and explore alternatives. I shall be considering Geach's claims solely as pertaining to the philosophy of language and philosophical logic, though much of the interest of the concept of relative identity concerns its applicability to other areas: the metaphysical controversy about *personal identity* and the debate in philosophical theology on the doctrine of the Trinity (see Geach, 1961; 1977; Cartwright, 1987; Van Inwagen, 1990; Cain, 1989). I shall set out Geach's views under six headings: (1) the non-existence of absolute identity, (2) the sortal relativity of identity, (3) the derelativization thesis, (4) the counting thesis, (5) the thesis of the irreducibility of restricted quantification, and (6) the 'name for an A' / 'name of an A' distinction.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

I shall examine the main tenets of Geach and his opponents with regard to (1), (2), and (3), which are the core of his position. I begin with thesis (1).

The Non-existence of Absolute Identity

On the classical view of identity it is an equivalence relation which satisfies Leibniz's Law. These formal properties ensure that within any theory expressible by means of a fixed stock of predicates, quantifiers, and truth-functional connectives, any predicates which can be regarded as expressing identity will be extensionally equivalent. But they do not ensure that a two-place predicate does express identity within a theory, for its descriptive resources may not be rich enough to distinguish items between which the equivalence relation expressed by the predicate holds (Geach, 1972, pp. 238–247).

Geach calls a two-place predicate which has these formal properties in some theory an 'I-predicate' (actually 'I-predicable') relative to that theory. Relative to another, richer, theory the same predicate, interpreted in the same way, may not be an 'I-predicate.' If so, it will not, and did not, even in the poorer theory, express identity.

However, Quine has suggested that when a predicate is an I-predicate in some theory only because the language in which the theory is expressed does not allow one to distinguish items between which it holds, one can reinterpret its sentences so that the I-predicate in the newly interpreted theory does express identity. Each sentence will have just the same truth-conditions under the new interpretation and the old. Thus Quine suggests that in a language in which persons of the same income are indistinguishable its predicates may be reinterpreted so that the predicate which previously expressed *having the same income* comes now to express identity. The universe of discourse now consists of income groups, not people. The extension of an n-place predicate is a class of n-member sequences of income groups (see Quine, 1963, pp. 65–79). Any two-place predicate expressing an equivalence relation could be an I-predicate relative to some theory, and Quine's suggestion will be applicable to any such predicate.

In his (1967; reprinted in his 1972, pp. 238–247) Geach objects to Quine that applying this procedure leads to a 'baroque Meinongian ontology' inconsistent with Quine's own expressed preference for 'desert landscapes' (1972, p. 245). He concludes that the only tenable position is his own, that identity is relative.

To understand this we can begin by considering the following passages. In the first, Geach, as he often does, compares and contrasts his position with that of Frege:

> When one says 'x is identical with y' this, I hold, is an incomplete expression. It is short for 'x is the same A as y,' where 'A' represents some count noun understood from the context of utterance – or else it is just a vague expression of a half-formed thought. Frege emphasized that 'x is *one*' is an incomplete way of saying 'x is one A, a single A' or else has no clear sense: since the connection of the concepts *one* and identity comes out just as much in the German 'ein und dasselbe' as in the English 'one and the same,' it has always surprised me that Frege did not similarly maintain the parallel doctrine of relativized identity, which I have just briefly stated. (1967, p. 3)

Geach often associates his thesis that identity is relative with the notion of a *criterion of identity*: "I maintain that it makes no sense to judge whether x and y are 'the same' or

whether x remains 'the same' unless we add or understand some general term 'same F.' That in accordance with which we thus judge as to the identity, I call a *criterion* of identity." And he takes his view to have the implication that "x is the same A as y" does not "split up" into "x is an A (and y is an A) and x is the same as … y" (1962, pp. 39, 152). (On criteria of identity, see further Chapter 39, OBJECTS AND CRITERIA OF IDENTITY.)

He also remarks (1980, p. 181), "On my own view of identity I could not object in principle to different As being one and the same B."

This last quotation gives us a way of understanding Geach. Say that an equivalence relation R is *absolute* if and only if, if x stands in it to y, there cannot be some other equivalence relation S, holding between anything and either x or y, but not holding between x and y. If an equivalence relation is not absolute it is *relative*. Now the question can be raised whether there are any *absolute* equivalence relations.

Geach's claim is that there are not. This is vaguely stated, however. Given the definition of an absolute equivalence relation above, classical identity must be an absolute equivalence relation if it exists, as must any necessarily uninstantiated equivalence relation. So, stated more precisely, Geach's claim is that *any expression for an absolute equivalence relation has the null class as its extension*. This entails that *there can be no expression for classical identity*, given that we understood it as the relation everything stands in to itself and nothing else. This is what Geach argues against Quine. We shall look at his argument later.

The Sortal Relativity of Identity

It may be that whenever 'A' is interpretable as a count noun or sortal term in a language L, the expression (interpretable as) 'x is the same A as y' in language L will be satisfied by a pair of things only if the *I-predicate* of L is satisfied by the pair. So no truth of the form 'x and y are the same A but different Bs' will be expressible in the language. Geach's contention that this is a possibility is thus an additional thesis – the *thesis of the sortal relativity of identity* – which is not entailed by his thesis of the non-existence of absolute identity, and in fact, and far more importantly, does not entail the latter, and so cannot just be rejected along with it. This thesis is the central point at issue between Geach and Wiggins. It entails that a relation expressible in the form 'x is the same A as y' in a language L need not entail indiscernibility even by the resources of L. Geach argues for it by illustrative examples – the cases of the cat on the mat (1980, p. 215), Heraclitus and the bath water (1962, pp. 150–151), and men and heralds (1980, pp. 174 ff.). We shall look at these later. In each case Geach argues that the relevant relation does not express indiscernibility even by the resources of our language.

The Derelativization Thesis

The sortal relativity thesis depends for its significance on the distinction between sortal (or what Geach calls 'substantival') terms and non-sortal (or 'adjectival') terms. For we can simply introduce *by abbreviative definition* an expression of the form 'x is the same A as y' to denote a relative equivalence relation R and, again by abbreviative definition, an expression of the form 'x is an A' to denote the property of being R to something or other, and then it may turn out that in the language thus expanded some statement of the form 'x is an A,

y is an A, x is the same A as y but x and y are different Bs' is true. This is how Geach does introduce the concept of a surman, defining 'x is the same surman as y' to mean 'x and y are men with the same surname' and then 'x is a surman' to mean the same as 'x is the same surman as something or other.' But his opponents will simply deny that 'surman,' so introduced, functions as a sortal term, that is, conveys a genuine criterion of identity. (See Chapter 39, OBJECTS AND CRITERIA OF IDENTITY.)

Geach's response to this line of objection is to offer his own account of the distinction between sortal terms and non-sortal terms – an account which is consistent with the thesis of the sortal relativity of identity.

The basic distinction is between those terms 'A' such that 'same A' makes sense, and those terms of which this is not true. In his (1962) and (1967) Geach, following Aquinas, calls this the distinction between 'substantival' and 'adjectival' terms. He illustrates it by reference to Frege's remarks about the number of red things:

> Frege said that only such concepts as "sharply delimited" what they applied to, so that it was not "arbitrarily divisible," could serve as units for counting … Frege cagily remarks that in other cases, e.g., "red things," no finite number was determined. But, of course, the trouble about counting the red things in a room is not that you cannot make an end of counting them, but that you cannot make a beginning, you never know whether you have counted one already, because "the same red thing" supplies no criterion of identity. (1962, p. 63)

Thus, according to Geach, 'red thing' is an adjectival term because 'same red thing' provides no criterion of identity and makes no sense, whereas 'apple,' say, and, as he goes on to mention, 'gold,' are substantival terms because 'same apple' and 'same gold' do provide criteria of identity (the difference between the latter two terms is the difference between *count nouns* and *mass terms*, see Chapter 39, OBJECTS AND CRITERIA OF IDENTITY).

But why does 'same red thing' not make sense, whereas 'same apple' does? And what is the relation in the latter case between the two-place predicate 'is the same apple as,' and the one-place predicate 'is an apple'?

Geach's proposes that the latter is derived from the former by what Quine has called *derelativization*. Quine writes: "commonly the key word of a relative term is also used *derelativized*, as an absolute term to this effect: it is true of anything x if and only if the relative term is true of x with respect to at least one thing. Thus anyone is a brother if and only if there is someone of whom he is a brother" (1960, p. 106). It would be nonsense to suppose that the explanation could go the other way round; that we could start with 'is a brother' and then go on to explain 'is a brother of.' Just so, Geach claims, with respect to 'is an apple' and 'same apple.' 'Is an apple' is definable by derelativization as 'is the same apple as something,' and 'the same' in 'the same apple as' is not a syntactically separable part but an index showing we have here a term for a relation with certain logical properties, like the 'of' in 'is a brother of' (1973, p. 291).

'A' is a substantival term, then, according to Geach, if 'is (an) A' is to be explained as formed by derelativization from 'is the same A as.' For, since 'the same' is merely an index of a certain type of relation, we cannot start with the monadic predicate 'is (an) A' and then explain the relational predicate 'is the same A as' in terms of it.

So, Geach claims, how he introduces talk of surmen is not a mere trick, but a faithful representation of the way in which substantival terms generally are introduced.

Of course, Geach need not hold that every equivalence relation can serve in this way to introduce a substantival term; in fact, in his later writing on identity (1991, pp. 294 ff.) he

denies this and disowns the 'surman' example. But his position is that every substantival term is derelativized from an expression for an equivalence relation and that merely *relative* equivalence relations may serve in this role.

To refute Geach, then, what his opponents must do is to point to features of the semantics of substantival terms which are incompatible with their being derelativized from expressions for relative equivalence relations, whilst to establish his thesis Geach must demonstrate that there are no such features.

It is in the light of this that Geach's views on *counting* must be understood.

The Counting Thesis

To count we distinguish items not yet counted from those already counted, and identify ones already counted as being among those already counted. Many philosophers think that if x is an A and y is an A and x and y are not (classically) identical then x and y cannot be counted as *one* A. Accordingly, when counting As one must count them as one only if they are identical. But, in fact, as Geach points out, it is perfectly possible to count by a relation weaker than classical identity – a relative equivalence relation. Suppose R is a relation weaker than identity which holds among As and which sorts the As into equivalence classes, then one can count As according to the rule that As x and y are to be counted as one just in case x bears R to y. To do so one assigns *one* to any A and to any A which bears R to that A, and to no other A; *two* to any A to which a number has not yet been assigned, to any A which bears R to it and to no other A, and so on. The final number reached will be the count of As when counting by R, and it may be smaller than the number arrived at when counting by classical identity.

It is, of course, a further question whether we ever do count by a relative equivalence relation, as Geach claims. But the correctness of Geach's counting thesis – the thesis that we *can* do so without falling into confusion or inconsistency – is enough to show that the mere fact that a noun is a count noun does not suffice to show that it *cannot* be understood as a derelativization of an expression for a relative equivalence relation. Moreover, given this analysis of counting, Geach can demand his opponents explain how, on their view, any term can be (logically) adjectival. For if the relation we count by is always identity then it is at first sight hard to see why the distinction between substantival and adjectival terms does not simply collapse; that is, it is hard to see how 'same A' can make any better sense in some cases (e.g., 'same man') than in others (e.g., 'same red thing'), for in every case 'x is the same A as y' means 'x is an A and y is an A and x and y are identical.'

The Irreducibility of Restricted Quantification

An important component of Geach's position is his thesis that for any sortal term 'A' there is a distinction between restricted quantification over As and unrestricted quantification over things that *are* As.

'Some man is F,' Geach holds, is not equivalent to 'something is a man and is F,' and 'every man is F' is not equivalent to 'everything, if it is a man, is F.' Again, he says, if 'A' and 'B' are two sortal terms, 'Every (some) A is F' need not be equivalent to 'Every (some) B is F,' even

if 'Every A is a B' and 'Every B is an A' are both true. Thus, for example, he claims, 'Every (some) man is F' need not be equivalent to 'Every (some) herald is F' even if every man is a herald and every herald is a man.

These claims about restricted quantification are, in fact, consequences of Geach's thesis of the sortal relativity of identity. We can see this by looking at two of the examples he uses to argue for the sortal relativity thesis.

First, the case of the cat on the mat. If Tibbles is on the mat and is the only cat there, there will, nonetheless, Geach claims, be many numerically distinct individuals on the mat which are cats and the same cat as Tibbles. For each proper part of Tibbles which is smaller than Tibbles by just one hair is a cat and (since there is only *one* cat on the mat) the same cat as Tibbles and every other such part. If so 'Some cat is F' is not equivalent to 'something is a cat and is F.' For, in this situation, 'some cat is F' is true just in case 'the cat on the mat is F' is true. But 'the cat on the mat lacks hair number one' (the first hair considered) is not true, so 'some cat lacks hair number one' is not true, but 'something is a cat and lacks hair number one' is.

Second, the case of Heraclitus and the river. If Heraclitus bathes twice in the river and, as Geach claims, the river is at any moment a collection of water molecules and so is the same water as the collection of water molecules then in the river bed (since there are not two collections of water molecules occupying exactly that space at that time), then it will be true that Heraclitus bathes in *something which is water* on two successive occasions, but it will be false (since new waters are ever flowing in) that there is *some water* that Heraclitus bathes in twice.

If Geach is right that identity is sortal relative, then, the irreducibility of restricted quantification follows. But one can accept the latter thesis without accepting the former.

The 'Name for an A'/'Name of an A' Distinction

The distinction between restricted and unrestricted quantification, if accepted, carries with it a distinction between two senses in which a name may name an A: a name may name *something which is an A*; more strongly, it may name *some A*. In the former case Geach calls it a name *of* an A; in the latter case a name *for* an A. Thus any non-empty name *for* an A is also a name *of* an A, but, if restricted quantification is irreducible, a name *of* an A need not be a name *for* an A. In the Tibbles case, for example, 'Tibbles' is both a name *for* a cat and a name *of* a cat, but if 'c' names a proper part of Tibbles which is a cat, it will be a name *of* a cat, but not a name *for* a cat.

With this distinction made Geach is able to explain the truth-conditions of statements containing restricted quantification and their relation to the truth-conditions of statements containing unrestricted quantification as follows:

'F (some A)' is true iff 'F(a)' is true for some interpretation of 'a' as a name *of* and *for* an A;
'F (any A)' is true iff 'F(a)' is true for any interpretation of 'a' as a name *of* and *for* an A.

If we delete from the above truth-conditions for 'F(some A)' and 'F(any A)' the restriction to proper names *of* and *for* an A we obtain truth-conditions for 'For some x, Fx' and 'For any x, Fx,' respectively. It is worth noting that Geach does *not* intend that these explanations should be read as employing substitutional quantification (1978).

Thus, we can say, a name 'a' which names something which is an A is a name *for* an 'A' if 'F(a)' is a sufficient condition for the truth of 'F(some A)' otherwise it is merely a name *of* an A.

(A complication to be mentioned here is that Geach holds that there is no absolute distinction between general and proper names, since a name may be at the same time a name of several As and a name of just one B. Consequently, Geach would say, in the Tibbles case 'Tibbles' is a proper name *for* and *of* a cat, but is also a general name *of* each proper part of Tibbles which qualifies as a cat. This position does not appear to be logically required by his other views. If 'Tibbles' names Tibbles, and Tibbles is the same cat as c (a proper part of Tibbles which qualifies as a cat), why must we infer that 'Tibbles' names c unless *same cat* is an absolute equivalence relation? But, by hypothesis, it is not. In the sequel, therefore, I will concentrate on Geach's views about identity and leave aside his views on general and proper names.)

This account of the name *for* an A/name *of* an A distinction takes for granted the distinction between restricted and unrestricted quantification. However, Geach thinks that the former distinction can be explained independently and thus can be used to cast light on the latter. A name *for* an A, Geach suggests, can be explained as a name associated with the criterion of identity: *same A*. A name of an A which is not a name *for* an A, on the other hand, is a name which names something which is an A but is not associated with that criterion of identity.

The idea of a criterion of identity is a much-stressed element of Geach's conceptual repertoire, which he derives from Frege and Wittgenstein, but it is an idea which he shares with many philosophers, including strong opponents of his views on relative identity. It is a standard, though not uncontroversial view, that reference is only possible against the background of a criterion of identity, and hence that any proper name must have a sense (not necessarily an individuating sense of the sort attacked by Kripke, 1980) which has a criterion of identity as a component (see Chapter 39, OBJECTS AND CRITERIA OF IDENTITY).

The general idea can be understood as follows. To assign a name a use is to determine its contribution to the truth-conditions of the sentences in which it occurs. To contend that to introduce a name one must associate it with a criterion of identity is, then, to say that one must do this to fix its contribution to the truth-conditions of sentences in which it occurs.

To consider why this is thought to be the case would take us too far afield. But for now it suffices to note that what is disputable about Geach's position is not the importance he gives to criteria of identity but his claim that a name *of* an A – a name which names something which is an A – can *fail* to be a name *for* an A and that something can be an A without satisfying the criterion of identity for As.

I now turn to an examination of the main arguments for and against Geach's claims.

Geach versus Quine: A Baroque Meinongian Ontology

Geach argues for his thesis (1), that absolute identity does not exist, by trying to show that absurdities result from Quine's claim that one can always reinterpret the range of the quantifiers in a language L so as to ensure that the I-predicate of L expresses absolute identity, and not merely indistinguishability by the stock of predicates contained in L. To be relevant to its target the argument must be read as assuming that *if* absolute identity is expressible in language at all *then* one can always reinterpret the range of the quantifiers in any language

L in such a way as to ensure that the I-predicate of L expresses absolute identity; but this assumption seems unexceptionable. Geach argues that this Quinean claim leads to a 'baroque Meinongian ontology.' There are, however, two versions of Geach's argument, an earlier one and a later one, and these need to be considered separately, since the earlier argument is vulnerable to a criticism which does not apply to the later one.

In its earlier version the argument goes as follows. Suppose we have a language L containing expressions for equivalence relations E_1, E_2, E_3 and a theory T expressible in L in which these expressions are employed. Then, for each such expression E_n we can consider that sub-language of L (L_n) in which that expression is an I-predicate, and that fragment of T (T_n) expressible in that sub-language. Adopting Quine's suggestion, we can then reconstrue the range of the quantifiers in each L_n and reinterpret the predicates of L_n so that while each true sentence of T which is also a sentence of T_n remains true in T_n, E_n in L_n no longer expresses a relation which holds between distinct items, but rather the relation of absolute identity. The range of the quantifiers in each L_n will now be different from their range in any other L_n, and also different from their range in L. For instance, if we start off with a language in which we quantify over token words, and in which the predicates 'is the same token word as,' 'is equiform to,' and 'has the same dictionary entry as' all occur, we may consider fragments of this language, and correspondingly fragments of theories expressible in this language, in which these various predicates qualify as I-predicates.

Following Quine's suggestion we may then reconstrue the quantifiers and reinterpret the predicates in these various language-fragments in a way that ensures, for example, that 'has the same dictionary entry as' expresses absolute identity in the language-fragment in which it is the I-predicate. One way of doing this is to regard the quantifiers in this language-fragment as ranging over classes of words which have the same dictionary entry. Similarly, one may regard the quantifiers in the language-fragment in which 'is equiform to' is the I-predicate as ranging over classes of equiform words. Since equiform words need not have the same dictionary entry, nor words with the same dictionary entry be equiform, the ranges of the quantifiers in these two language-fragments will now be different, and different again from the range of the quantifiers in the original language, in which neither 'has the same dictionary entry as' nor 'is equiform to' is an I-predicate.

Now Geach does not claim, in the earlier version of his argument, that interpreting quantifiers in this way, so as to get at a relation of absolute identity, involves one in logical incoherences or absurdities, merely that it sins against a highly intuitive methodological program enunciated by Quine himself, namely that "as our knowledge expands we should unhesitatingly expand our ideology, our stock of predicables, but should be much more wary about altering our ontology, the interpretation of our bound name variables" (Geach, 1972, p. 243), and that it has a consequence possibly unwelcome to a lover of desert landscapes, namely that

> since a rich language L may allow for our carving many sub-languages, L_1, L_2, L_3 … out of it, users of L are committed to the existence, not only of a realm of objects for which the I-predicable of L itself gives the criterion of absolute identity, but also for each of these possible sub-languages L_n, of a distinct realm of objects for which the I-predicable of L_n gives the criterion of absolute identity. (1972, p. 248)

Geach's argument is thus that in view of the mere *possibility* of carving L_1, L_2, L_3 … out of L, if the thesis maintained by Quine is right, users of L will be ontologically committed to

any number of entities which are not spoken of, or quantified over, in L. They will be so committed because any sentence of L which is also a sentence of some sub-language L_n will have just the same truth-conditions in L and L_n and hence also in any theory T expressible in L and any theory got from T by mere omission of the sentences of L which are not sentences of L_n, but "[it] is, of course, flatly inconsistent to say that as a member of a larger theory a sentence retains its truth-conditions but not its ontological commitment" (Geach, 1973, p. 299).

The crucial premise of this argument, it therefore emerges, is that sameness of truth-conditions entails sameness of ontological commitment. But, however it may be with other notions of ontological commitment, this is not true of Quine's. For Quine, the ontological commitments of a theory are those entities which must *lie within the domain of quantification* of the theory if the theory is to be true; or, alternatively expressed, those entities the predicates of the theory have to be true of if the theory is to be true. A theory is not, if I may so express it, ontologically committed to what is required to be in *the universe* if it is to be true, but merely to what it is required to be in *its universe* if it is to be true. Because this is so there is no argument from sameness of truth-conditions to sameness of ontological commitments.

Thus, as an *ad hominem* argument against Quine (which is how he himself describes it) Geach's argument, in the earlier version now being discussed, has to be judged a failure.

Matters stand differently with the later version of the argument, though it, too, in the end turns out not to be cogent (the criticism following is indebted to Dummett, 1991). The difference between the earlier and later version is that in the later (to be found in Geach, 1973) Geach's claim is not merely that Quine's thesis about the interpretation of quantification has a consequence which is unpalatable and "possibly unwelcome to a lover of desert landscapes," but that it leads to an out-and-out logical absurdity, the existence of *absolute* surmen. Because Geach is now making this stronger claim, the objection that his argument depends upon the incorrect assumption that sameness of truth-conditions entails sameness of ontological commitments is no longer relevant. In order to make out his case Geach has to establish just two points. First, that there are sentences of English (supplemented by the predicate 'is the same surman as') which are evidently true and which, considered as sentences of that fragment of English in which 'is the same surman as' is an I-predicate, when this is interpreted in the way Quine suggests, can be true only if absolute surmen exist. And second, that the existence of absolute surmen (entities for which 'is the same surman as' expresses absolute identity) is absurd.

But in the end Geach fails to establish these points. Quine would say that, for the fragment of English in question, the domain of the variables can be considered as consisting of classes of men with the same surname and the predicates interpreted as holding of such classes. Thus, the predicate 'is the same surman as' will no longer be true of *men* if we adopt Quine's suggestion (I am writing, remember, in English, not in the fragment of English under discussion), but rather of classes of men with the same surname – these, then, will be the entities which are Geach's 'absolute surmen.' Now, Geach attempts to rule out such a suggestion by the argument that 'Whatever is a surman is by definition a man.' But this argument fails. The predicate 'is a man' will also be in the language-fragment in which 'is the same surman as' is the I-predicate; and so it, too, will be reinterpreted, if we follow Quine's suggestion, as holding of classes of men with the same surname. Thus the sentence 'Whatever is a surman is a man' will be true in the language-fragment interpreted in Quine's way, just as it is in English as a whole. What will *not* be true, however, is that whatever the

predicate 'is a surman' is true of, *as it occurs in the language-fragment reinterpreted in Quine's way*, is a thing of which 'is a man,' *as it occurs in English as a whole*, is true of. But Geach has no right to demand that this should be the case. Even so, this demand can in fact be met. For the domain of the interpretation of the language-fragment in which 'is the same surman as' in the I-predicate can be taken to consist of men, namely to be a class containing exactly one representative man for each class of men with the same surname. Thus, as Geach says, "absolute surmen will be just some among men" (1973, p. 300). Geach goes on, "There will, for example, be just one surman with the surname 'Jones'; but if this is an absolute surman, and he is a certain man, then which of the Jones boys is he?" But this question, which is, of course, only answerable using predicates which belong to the part of English not included in the language-fragment in which 'is the same surman as' is the I-predicate, is not an impossible one to answer. It is merely that the answer will depend upon the particular interpretation which the language-fragment has been given. Geach is, therefore, not entitled to go on "Surely we have run into absurdity." It thus seems that his argument for the non-existence of absolute identity fails.

Cats, Rivers, and Heralds

Geach's thesis (2) – his sortal relativity thesis – is, however, another matter. For it is neither entailed by, nor entails, the thesis of the non-existence of absolute identity. Geach argues for it by appeal to well-known examples: the case of the cat on the mat, Heraclitus and the river, and the men and heralds case. I shall concentrate on the case of the cat on the mat.

There are two versions of the argument about the cat on the mat. One version goes like this (see Wiggins, 1968). Suppose Tibbles, is on the mat. Now consider that portion of Tibbles which includes everything except the tail, call it 'Tib.' Since Tibbles and Tib are different sizes they are non-identical. But if we amputate the tail they now occupy exactly the same space. If Tibbles is still a cat, it is hard to see why Tib is not, even if it was not before. Yet they are distinct individuals, because their histories are different. But there is just *one* cat present. So they cannot be distinct *cats*. They must be the same cat, even though distinct individuals; and so being one and the same cat must be a relative identity relation, that is, a relation which does not ensure the indiscernibility of its terms.

A second version (Geach, 1980) goes as follows. Tibbles is sitting on the mat and is the only cat sitting on the mat. But Tibbles has at least 1,001 hairs. Geach continues:

> Now let c be the largest continuous mass of feline tissue on the mat. Then for any of our 1,000 hairs, say h_n, there is a proper part c_n of c which contains precisely all of c except that hair h_n: and every such part c_n differs in a describable way both from any other such part say c_m, and from c as a whole. Moreover, fuzzy as the concept *cat* may be, it is clear that not only is c a cat, but also any part c_n is a cat: c_n would clearly be a cat were the hair h_n to be plucked out, and we cannot reasonably suppose that plucking out a hair *generates* a cat, so c_n must already have been a cat. (1980, pp. 215–216)

The conclusion, of course, is the same as before: all the distinct entities must qualify as cats since there is only one cat on the mat and *same cat* must be a merely relative identity relation.

The second version is vulnerable to an objection – that the concept of *cat* satisfies a maximality requirement: that nothing can be *both* a proper part of a cat *and* a cat – which

does not apply to the first. But obviously neither version will convince an opponent, who will simply deny that Tib or any of the entities distinct from Tibbles in the situation is ever predicatively a cat, pointing to the modal and historical properties possessed by Tibbles not possessed by the other entities in support.

However, a defender of Geach's position will want to know the relevance of the differences. They prove that Tib and the like are numerically distinct from Tibbles, but we knew that all along. The debate is about whether they are cats. Are they not too cat-like not to be? How can things so similar to a paradigm cat not be cats? And in response to the denial that Tib is a cat after the operation in the version of the argument given by Wiggins, the relative identity theorist can ask why it should not be a cat afterwards, even granted that it is not before. Granted it has an unfeline past, why does that exclude its having a feline present (and future)? Unless it is simply stipulated that however cat-like something is at a time it cannot qualify as a cat unless it has the right sort of past (and future?).

In fact it is clear that there are three possible lines of solution to the puzzle of the cat on the mat:

1. Geach is wrong. The correct definition of 'cat' applies only to Tibbles itself and not to Tib and its like. However, it is correct to say that Tib is a cat, after the amputation, if not before, since 'is' has, as well as its predicative sense, various constitutive senses (Wiggins, 1968; Shoemaker, 1970; Lowe, 1989) and since Tib shares all its matter with Tibbles after the amputation it 'is,' then, in one legitimate sense, a cat.
2. Geach is right that many distinct individuals are present each of which is predicatively a cat. Counting is not always by identity. In counting cats we count by a weaker equivalence relation. Consequently, we speak correctly in saying that there is just one cat present, even though many distinct individuals are present each of which is predicatively a cat. But in counting cats the relation we count must be the one we express by 'is the same cat as.' This, then, must be the relevant relation. Thus 'is the same cat as' is an expression for a relative equivalence relation.
3. Geach is right that many distinct individuals are present, each of which is predicatively a cat, but counting is by identity. When counting cats, not everything that qualifies as a cat counts. 'There is just one cat on the mat' means 'some cat is on the mat and every cat which is on the mat is identical with that one.' The only entities to be counted when counting cats are those which fall within the range of the natural-language quantifying expressions, 'some cat' and 'every cat.' But it is only if 'some cat is F' is equivalent to 'something is a cat and is F' and 'every cat is F' is equivalent to 'everything, if it is a cat, is F' that these quantifying expressions must be taken to range over everything which qualifies as a cat. These equivalences do not hold. Of the 1,001 items in the situation which qualify as predicatively cats, only one – Tibbles – falls within the range of 'some cat' and 'every cat.'

It seems that the linguistic facts are consistent with each of these solutions, so the puzzle cannot count *decisively* either for or against Geach's view. But the availability of the third line of solution also makes it evident that no example of this type could even provide a *reason* for embracing the sortal relativity thesis, since the distinction between restricted and unrestricted qualification which is all that the type (3) solution relies upon is something to which a proponent of a Geachian type (2) solution is already committed. On grounds of economy, then, it seems that type (3) solutions to problems of the sort Geach describes are always preferable to type (2) solutions.

The availability of the type (3) solution also puts the position of the proponent of the type (1) solution in a clearer light. It makes it clear that opposing the concept of relative identity, by itself, provides no motive for insisting on the type (1) solution or for endorsing the 'is' of constitution. Such a motive requires an argument for rejecting the thesis of the irreducibility of restricted (sortal) qualification to unrestricted quantification, but it is hard to see what form such an argument might take. Since both sides agree that it is correct to say of something numerically distinct from Tibbles in the situation that it 'is' a cat but disagree over the meaning of 'is,' it appears that the crucial point at issue between the proponent of the type (1) solution and the proponent of the type (3) solution is whether '*is a cat*,' *understood as a syntactically simple predicate in which the 'is' is merely the 'is' of predication – a mere fragment of a predicate which expresses no property or relation by itself – applies univocally both to Tibbles and to (at least one of) the entities present in the situation described which are distinct from Tibbles*. But how this issue might be decided is wholly unclear.

However, there is a third variant of the puzzle which brings new considerations into play. This is Lewis's version (Lewis, 1993).

Lewis presents the puzzle as one about vagueness. He then offers two solutions, in accordance with his own preferred account of vagueness as always having its source in "semantic indecision," the view that the reason why 'the Outback' is vague is that no one has been fool enough to decide where it should be deemed to lie. There is not something out there in the world with imprecise boundaries. The only intelligible account of vagueness locates it in our thought and language (Lewis, 1986, p. 213).

The first of his solutions rest on a supervaluational semantics according to which what is definitely true is what remains true under all precisifications. It accounts for our confidence that there is just one cat on the mat consistently with the thesis that counting is always by identity. According to this solution there is no one thing on the mat which is clearly a cat, though it is clear that there is a cat and exactly one on the mat.

Lewis's second solution does not appeal to supervaluationism to deliver the verdict that there is just one cat on the mat. According to this solution there are many, numerically distinct, entities on the mat which are clearly cats, just as Geach says, but 'there is just one cat on the mat' is nonetheless true since all of these massively overlap – they are almost identical – and our statement 'there is just one cat on the mat' records the result of counting not by identity but by almost-identity. This solution does not require Geach's rejection of absolute identity, and so Lewis does not think of it as a relative identity solution. But it entails that where there is just one cat there may be many numerically distinct individuals, each of which is a cat, and counting is not always by identity, which is what Geach wishes to establish with the tale of the cat on the mat.

Here is Lewis's statement of the puzzle:

> Cat Tibbles is alone on the mat. Tibbles has hairs h_1, h_2, …, h_{1000}. Let c be all of Tibbles, and c_1 all of Tibbles except for h_1 and similarly for c_2, …, c_{1000}. Each of these is a cat. … Why? … [S]uppose it is spring, and Tibbles is shedding. When a cat sheds the hairs do not come popping off, they become gradually looser. By the end … the loose hairs are no longer parts of the cat. Sometime before the end, they are questionable parts…. Suppose each of h_1, h_2, …, h_{1000} is at this questionable stage. Now indeed all of c_1, c_2, …, c_{1000} and also c have equal claim to be a cat. (Lewis, 1993, p. 166)

Lewis's first solution appeals directly to the idea of vagueness as semantic indecision concerning the extension of 'cat.' Our concept of a cat is vague because we have not made the

term precise. But consistently with what has been established it means we can decide to make it entirely precise in any of a number of ways. Subsequent to each such decision, each such precisification, it will turn out that there is just one cat on the mat – the cat-candidates will be reduced to one. But each precisification will identify a different candidate as the cat on the mat. Comparably, 'orange' is vague. We have not decided the precise boundary between orange and yellow or orange and red. We can make it precise in numerous ways consistently with the clear truths about orange, for example, that whatever is orange is not red and that whatever is orange is colored. On each of these precisifications there will be a line which is the boundary between orange and red, but different precisifications will identify different lines as the boundary.

The difficulty with this first, 'one cat,' solution, this solution by disqualification, as Lewis calls it, is that it cannot be that on each precisification just one cat-candidate turns out to be a cat. Precisifying decisions must preserve clear truth. So one cannot decide to precisify 'cat' in such a way that c_{59} is a cat and c_{60} is not if the only difference between these is that one is on my left as I face the mat and the other not, or one is closer to an edge of the mat than the other, or one is slightly browner than the other. It is clearly true that such differences cannot make a difference in respect of cathood so a precisification is not eligible that makes them do so. Since amongst c and $c_1, \ldots, c_{1000}$ there will be pairs so related that there is no difference between them that could make a difference in respect of cathood, given the already established meaning of 'cat,' no decision could be made, consistently with the meaning 'cat' actually has, which determines that exactly one of the many cat-candidates falls within the (new) extension of 'cat.'

So, given the Lewisean assumption that semantic indecision is the source of all vagueness, we cannot say, consistently with holding it to be clearly true that there is one cat on the mat, that the cat-candidates are equally good, because they are all equally indeterminate examples of cathood.

Rather, we must say that they are all equal because they are all clearly cats. This is Lewis's second proposal. But, how, then, can there be just one cat on the mat?

Lewis's answer is that any two cat-candidates are almost completely identical, they massively overlap. And it is this relation we count by in counting cats (in almost every conversational context). So in almost every conversational context it is correct to say that there is just one cat on the mat.

But how then are we to understand the description '*the* cat on the mat'? We want it to be clearly true that the cat on the mat is hairy, but not clearly true that the cat on the mat includes hair number 17. Suppose we subject the description to Russellian translation. We then get as translations of 'the cat on the mat includes h_{17} (is hairy)':

> There is something that is identical to everything which is a cat on the mat and nothing else, and that includes h_{17} (is hairy)
> There is something that is identical to everything which is a cat on the mat and nothing else, and everything which is a cat on the mat includes h_{17} (is hairy)
> But these are all false, because nothing is identical to everything which is a cat on the mat.

But we can relax the translations by replacing 'identical' with 'almost identical.' Then both the translations' parts are no longer equivalent.

> There is something that is almost-identical to everything which is a cat on the mat and nothing else, and that includes h_{17}

is true.

> There is something that is almost-identical to everything which is a cat on the mat and nothing else, and everything which is a cat on the mat includes h_{17}

is false.

But both:

> There is something that is almost-identical to everything which is a cat on the mat and nothing else, and that is hairy

and:

> There is something that is almost-identical to everything which is a cat on the mat and nothing else, and everything which is a cat on the mat is hairy

are true.

So if 'the cat on the mat is F' is interpreted as receiving the truth-value given by the two relaxed Russellian translations if there is one, and as indeterminate otherwise, we get what we want, that 'the cat on the mat is hairy' is clearly true, and 'the cat on the mat includes hair 17' is not.

We also want 'Tibbles is hairy' to be clearly true and 'Tibbles includes hair 17' not to be, and, of course, we want 'Tibbles is the cat on the mat' to be clearly true. But this is secured immediately if the reference of 'Tibbles' is fixed via the description 'the cat on the mat,' which must at least imply that the truth of the identity is stipulated. Thus we get the Geachian distinction between names like 'Tibbles,' which are names for cats because their reference is stipulated by reference-fixing descriptions of the type 'the cat …' which are to be understood via relaxed Russellian translations, and names like 'c_{17},' which are merely names of cats.

This distinction also brings with it Geach's distinction between restricted (sortal) quantification and unrestricted quantification. In this context 'some cat is F' should be true if and only if 'the cat on the mat is F' is true, if and only if 'Tibbles is F' is true. So 'Some cat includes hair number 17' should not be true, though 'something is a cat and includes hair number 17' is true. Hence, in general, 'some cat is F' should be true iff 'a is F' is true for some interpretation of 'a' as a name of and for a cat.

So the verdict on Geach's thesis of sortal relativity must be tentative. It is not proven; but there seems to be no argument weighing conclusively against it. There seems no evident way to resolve the disputes over Geach's versions of the puzzle about the cat on the mat, whilst the Lewisean version, even if the considerations just gone through are accepted, provides support for it only given the substantial assumption that all vagueness must be identified with semantic indecision.

Substantival Terms and the Derelativization Thesis

This is not the case, however, with Geach's derelativization thesis: that every substantival term is to be explained as the derelativization of an expression for an equivalence relation. In this case it seems clear that Geach's contention is overambitious, as Dummett (see 1981; 1991) demonstrates.

The first counter-examples to the derelativization thesis Dummett notes are *derivative* count nouns, where a count noun 'A' is a derivative count noun when there is some count noun 'B' such that 'is the same A as' may be satisfactorily explained as 'is an A and is the same B as.'

As Dummett points out, ironically enough Geach himself draws attention to counter-examples of this class when he introduces the derelativization thesis. If Geach is right that 'is a brother' is derived from 'is a brother of,' it cannot also be understood as derived from 'is the same brother as.' Rather, we have to understand the latter as derived from 'is a brother' (or else, implausibly, reject it as meaningless); and the evident explanation is that 'is the same brother as' means 'is a brother and is the same man as.'

As Dummett argues, such nouns as 'postman' and 'baker' also seem to be exceptions. We understand 'is the same postman as' as meaning 'is a mail deliverer and is the same man as.' We do not have to learn 'is the same postman as' before we understand 'is a postman' and we cannot be thought of as required to derive the latter from the former by derelativization.

As Dummett notes, abstract nouns are also counter-examples to the derelativization thesis.

Consider the noun 'shape.' This is certainly a count noun, but it seems clear that Geach's derelativization thesis does not give its semantics. There is a competing account which is far more plausible, namely the account sketched out by Frege in the *Grundlagen*, using the concept of direction as his model. According to this account, the noun 'shape' may be thought of as introduced into the language as follows: we begin by introducing an expression 'has the same shape as' for an equivalence relation between material objects; we then introduce the functional expression 'the shape of,' explained in such a way as to yield the equivalence of 'the shape of x is the same as the shape of y' and 'x has the same shape as y'; and finally we explain 'x is a shape' to mean 'for some y, x is the shape of y.' We can supplement this account by stipulating that 'x is the same shape as y' is to mean 'x is a shape and x is the same as y,' which is equivalent to 'for some z, for some u, x is the shape of z and y is the shape of u, and z has the same shape as u.' This account seems superior to Geach's because it reflects the necessary order of language acquisition. There *could not* be a language in which it was possible to make reference to shapes but which did not contain a functional expression with the sense of 'the shape of.' This is because shapes, unlike, say, colors, are not possible objects of ostension: even against the background of an appropriate criterion of identity one cannot pick out a shape by pointing and saying 'this.' The only way to refer to a shape is as the shape of some already-identified object or region. Thus, a language could not contain the predicates 'is a shape' and 'is the same shape as' unless it also contained the functional expression 'the shape of,' and the Fregean account is in accord with this fact.

Another example for which the Fregean account seems plausible is 'nationality.' Here, too, it seems that we understand 'x is a nationality' and 'x is the same nationality as y,' respectively, to mean 'for some x, x is the nationality of y' and 'for some z, for some u, x is the nationality of z and y is the nationality of u and z has the same nationality as u,' understanding 'the nationality of' in such a way as ensures the equivalence of 'the nationality of x is the same as the nationality of y' and 'x has the same nationality as y,' that is, 'x is a citizen of the same country as y.' But the reason seems slightly different from the reason in the previous case. For, while shapes are not possible objects of ostension, nationalities are. If I point towards a man and say 'this nationality,' there may well be no choice, given the criterion of identity invoked, of objects to which I can be referring. But if I point and say 'this shape' there will always be more than one (if there is even one) possible object of reference compatible with the criterion of identity I have invoked. Nevertheless, it does seem that a language could not

contain any means of making reference to nationalities, and hence could not contain the predicates 'is a nationality' and 'is the same nationality as,' unless it also contained a functional expression with the sense of 'the nationality of.' This is because no one could understand the notion of 'a nationality' without being aware of those relations among human beings in which there being such things as nationalities consists: and he could not be aware of these without being able to refer to individual human beings and their nationalities.

Once we recognize, with Dummett, that abstract nouns are counter-examples to Geach's derelativization thesis, it becomes plausible that mass terms are also.

Consider the mass noun 'gold.' Like shape, and unlike colors or nationalities, parcels of gold are not possible objects of ostension. Pointing and saying 'this gold' does not determine which object I am referring to. This is because any proper part of a parcel of gold is itself a parcel of gold, but a distinct parcel from that of which it is a proper part. Thus, just as in order to identify shapes we must relate them to some other, already-identified objects or regions, as the shapes *of* those objects or regions, so, to identify a parcel of gold, one must relate it to some already identified object as the gold *of* that object; just as one may identify a shape as the *shape* of so-and-so's wedding ring, so one may identify a parcel of gold as the *gold* of her wedding ring; and, as the possibility of reference to shapes depends upon the existence of such means of identification, the same holds of the possibility of reference to parcels of gold. And so a language *could not* contain the means of making reference to parcels of gold, and hence could not contain the predicates 'is gold' and 'is the same gold as' (understood as applicable to parcels of gold), unless it contained a functional expression with the sense of 'the gold of,' as it occurs in 'the gold of her wedding ring.'

But in the light of this, the Fregean pattern of explanation seems to have as much plausibility for 'gold' as it has for 'shape' or 'nationality.' The predicate 'x is gold' is to be understood as meaning 'for some y, x is the gold of y,' and 'x is the same gold as y' as 'for some z, for some u, x is the gold of z and y is the gold of u and z is constituted of the same gold as u,' understanding 'the gold of' in such a way as ensures the equivalence of 'the gold of x is the same as the gold of y' and 'x is constituted of the same gold as y,' where 'is constituted of the same gold as' expresses an epistemologically prior relation in the same way as do 'has the same shape as' and 'has the same nationality as.'

If these suggestions are correct, Geach's derelativization thesis is far too ambitious: there are many substantival terms which are counter-examples. It does not follow, of course, that there are no substantival terms to which it *does* apply. And, in fact, it might seem that it must apply to what, following Dummett, can be called 'basic count nouns'; that is, substantival terms which are (a) not abstract nouns (like 'nationality' and 'shape'), (b) not mass nouns (like 'gold'), and (c) not derivative count nouns (like 'father' or 'postman'). For in the case of such basic count nouns it seems that the association with a criterion of identity which is definitive of a substantival term can be made in no other way: it cannot, as in the case of derivative count nouns, be derived from an association with a second count noun in terms of which the first is defined; nor can it, as in the case of abstract nouns and mass terms, be made in the way the Fregean pattern suggests, which requires that the associated criterion of identity be an equivalence relation between objects *other than* those to which the count noun applies.

However, once again Dummett suggests an alternative pattern of explanation. In the case of basic count nouns, he suggests (1981; 1991) the crucial point to recognize is that the associated criterion of identity is not an equivalence *relation* at all (where a relation is thought of as holding between *objects*).

We cannot give a correct representation of that level of our language at which we quantify over and refer to objects, Dummett thinks, unless we recognize a lower level at which no reference to or quantification over objects exists; formalized languages serve only to regiment the higher level. At the lower level, what takes the place of the use, at the higher level, of proper names and other singular terms to refer to objects is the use of demonstrative pronouns in what Dummett calls 'crude predications.' The distinctive feature of this use of demonstratives is that no criterion of identity has to be invoked to make their utterance understood; no answer to the question 'This what?' need be available. In such crude predications the predicate cannot, therefore, be one applicable to an object, but must be one expressing what Strawson has called a 'feature-placing concept.' Examples of such crude predications are 'This is sticky,' 'This is red,' and 'This is smooth.'

The transition to the higher level, at which reference to and quantification over objects takes place, Dummett suggests, is mediated by what he calls 'statements of identification,' that is, statements of the form, 'This is the same X as that' where 'X' is a basic count noun. A child does not actually acquire the word 'cat' in the first place by learning to point simultaneously to, say, the head and tail of a cat, and to say, 'That is the same cat as that.' But this, nevertheless, correctly represents what is involved in the move from the lower level of language to the higher, namely the acquisition of a criterion of identity by which we can determine where one cat leaves off and another begins.

But a statement of identification, like a crude predication, does not *itself* involve any reference to objects, since in itself it is merely a crude relational statement like 'This is darker than that.' Hence the criterion of identity associated with a basic count noun is not an equivalence relation between objects, either objects of the sort to which the count noun applies, or objects of another sort: '… is the same X as …,' as used in statements of identification, is *like* an expression for an equivalence relation, but it does not stand for such a relation, since it is not, at this stage, used to express a relation between *objects* at all. To grasp the criterion of identity associated with a basic count noun, it is thus not necessary to have any prior conception of objects of any sort. To think otherwise, Dummett suggests, is Geach's basic mistake.

These suggestions of Dummett's seem entirely correct. But it is important also to see the extent of the agreement between Dummett and Geach. The main emphasis of Geach's work on identity has always been on the unusability of an absolute identity relation as a *criterion* of identity. It is customary in the literature (see Chapter 39, OBJECTS AND CRITERIA OF IDENTITY, §5) to distinguish two forms of identity criteria, one-level and two-level. A two-level criterion of identity specifies the criterial relation as a relation holding between entities other than those for which the criterion is being given. Frege's criterion of identity for directions or the Fregean criterion of identity for shapes just explained are examples. A one-level criterion of identity states the criterion of identity for a type of object as a relation between objects of that very type: thus it says, for example, that classes are the same if and only if they have the same members or events are the same iff they have the same causes and effects. In a two-level criterion the criterial relation specified is a relative equivalence relation. And Dummett in effect repudiates the legitimacy of one-level criteria. According to his suggestions, the criterion of identity associated with a substantival term must either be given by an expression for a relative equivalence relation not holding between the objects to which the general term applies, or it must be given by an expression which does not designate a relation, *a fortiori*, not an absolute equivalence relation, at all.

A one-level criterion of identity, of the type Dummett implicitly repudiates as a legitimate criterion of identity, takes the form:

If x is an A and y is an A then x = y if and only if Rxy.

This is equivalent to the conjunction of:

(1) If x is an A then Rxx

and

(2) If x is an A, everything which is an A and R-related to it is identical to x.

Thus it amounts to:

(3) Anything which is an A is the A which is R-related to it.

In this the definite description must be understood in Russell's original fashion. So this gives a necessary condition for being an A by way of giving a criterion of identification for every A.

There are thus three points to be made about such one-level 'criteria of identity,' which confirm Dummett's repudiation of them as properly speaking criteria of identity at all. First, they merely give necessary conditions for falling under the sortal concepts they inform us about, that is, they contribute to the specification of the extensions of the sortal predicates, they do not specify (absolute) identity conditions in addition, or tell us what identity consists in for things of the sort. Second, one cannot infer from the fact that the one-level criterion of identity for As does not guarantee identity for Bs that As are not Bs. It does not follow from the fact that the one-level criterion of identity for classes, for example, does not guarantee property identity that classes are not properties (Cartwright, 1967; Geach, 1972). Third, it is only because the criterial relation specified is a relative equivalence relation that a one-level 'criterion of identity' for things of a sort says anything that is not true equally of things of every sort. More precisely, it is only because the relation specified is a relative equivalence relation that (2) adds anything to (1), in which no expression for identity occurs at all. Thus while classes are the same iff they have the same members, and events perhaps the same if and only if they have the same causes and effects, the thought that these are 'criteria of absolute identity' of a type incompatible with Geach's central thesis of the sortal relativity of identity is mistaken.

If Dummett is correct, Geach is right to deny that an absolute equivalence relation can serve as a criterion of identity: the criterion of identity associated with a general term (and hence, derivatively, with a proper name) must either be given by an expression for a relative equivalence relation not holding between the objects to which the general term applies (as in the case of abstract nouns and mass nouns), or it must be given by an expression which does not designate a relation between objects at all, and *a fortiori* does not designate an absolute equivalence relation. This, I would suggest, is the most important lesson to learn from Geach's work.

References

Cain. J. 1989. "The doctrine of the Trinity and the logic of relative identity." *Religious Studies*, 25(2): 141–152.

Cartwright, R. 1967. "Classes and attributes." *Noûs*, 1(3): 231–241.

Cartwright, R. 1987. "On the logical problem of the Trinity." In *Philosophical Essays*. Cambridge, MA: MIT Press.

Chisholm. R. M. 1969. "The loose and popular and strict and philosophical senses of identity." In *Perception and Personal Identity*, edited by N. Care and R. H. Grim. Cleveland, OH: Ohio University Press.

Chisholm. R. M. 1970: "Identity through time." In *Language, Belief and Metaphysics*, edited by H. E. Kiefer and M. K. Munitz, pp. 163–182. Albany, NY: State University of New York Press.

Chisholm. R. M. 1976. *Person and Object*. London: Allen and Unwin.

Dummett. M. 1973. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth, 1981.

Dummett. M. 1981. *The Interpretation of Frege's Philosophy*. Cambridge, MA: Harvard University Press.

Dummett. M. 1991. "Does quantification involve identity?" In *Peter Geach: Philosophical Encounters*, edited by H. A. Lewis, pp. 161–184. Dordrecht, Netherlands: Kluwer Academic.

Geach. P. T. (with G. E. M. Anscombe). 1961. *Three Philosophers*. Ithaca, NY, and London: Cornell University Press.

Geach. P. T. 1962. *Reference and Generality*, 3rd edn. Ithaca, NY: Cornell University Press, 1980.

Geach. P. T. 1967. "Identity." *Review of Metaphysics*, 21(1): 3–12. Reprinted in Geach, 1972, pp. 238–247.

Geach. P. T. 1972. *Logic Matters*. Blackwell. Oxford.

Geach. P. T. 1973. "Ontological relativity and relative identity." In *Logic and Ontology*, edited by M. K. Munitz, pp. 287–302. New York: New York University Press.

Geach. P. T. 1977. *The Virtues: The Stanton Lectures 1973–74*. Cambridge: Cambridge University Press.

Geach. P. T. 1978. "Evans on quantifiers." *Canadian Journal of Philosophy*, 8(2): 375–378.

Geach. P. T. 1980. *Reference and Generality*, 3rd edn. Ithaca, NY: Cornell University Press.

Geach. P. T. 1991. "Replies." In *Peter Geach: Philosophical Encounters*, edited by H. A. Lewis. Dordrecht, Netherlands: Kluwer Academic.

Kripke. S. 1980. *Naming and Necessity*. Oxford: Blackwell.

Lewis. D. K. 1976. "Survival and identity." In *The Identities of Persons*, edited by A. Rorty. Berkeley: University of California Press, pp. 17–40. Reprinted in *Philosophical Papers*, vol. 1. Oxford: Oxford University Press 1983, pp. 55–73.

Lewis. D. K. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis. D. K. 1993. "Many, but almost one." In *Ontology, Causality, and Mind: Essays on the Philosophy of D. M. Armstrong*, edited by K. Campbell, J. Bacon, and L. Reinhardt, pp. 23–38. Cambridge: Cambridge University Press.

Lowe, E. J. 1989. *Kinds of Being*. Oxford: Blackwell.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1963. *From a Logical Point of View*. New York: Harper and Row.

Quine, W. V. O. 1973. *The Roots of Reference*. La Salle, IL: Open Court.

Shoemaker, S. 1970. "Wiggins on identity." *Philosophical Review*, 79(4): 529–544.

Van Inwagen, P. 1990. "And yet they are not three gods but one god." In *Philosophy and Christian Faith*, edited by T. Morris. Notre Dame, IN: University of Notre Dame Press.

Wiggins. D. 1967. *Identity and Spatio-temporal Continuity*. Oxford: Blackwell.

Wiggins. D. 1968. "On being in the same place at the same time." *Philosophical Review*, 77(1): 90–95.

Wiggins. D. 1980. *Sameness and Substance*. Oxford: Blackwell.

Further Reading

Geach. P. T. 1975. "Names and identity." In *Mind and Language: Wolfson College Lectures*, edited by S. Guttenplan, pp. 139–158. Oxford: Clarendon Press.

Geach. P. T. 1979. "Existential or particular quantifier?" In *Ontology and Logic* (Proceedings of an International Colloquium. Salzburg, September 21–24, 1976), edited by P. Weingartner and E. Mascher. Berlin: Duncker and Humblot.

41

De Jure Codesignation

JAMES PRYOR

This chapter surveys a novel kind of semantic structure that has been posited by Mark Richard, Kit Fine, Ángel Pinillos, and others. Their commitments will be explained as we proceed. I discuss four potential areas of application:

- §1 focuses on anaphora, especially cases that can't be handled by a "bound-variable" analysis ("strict" and donkey sentences). I also distinguish our target view from other treatments of anaphora.
- §§2 through 6 discuss the semantics of attitude reports.
- §6 also contrasts two kinds of complex anaphoric dependency. (This may overlap in part or whole with the phenomena mentioned in §1.)
- §7 explains a difference in how functions in a programming language can be sensitive to the identity of their operands.

1

Our stage is set by a range of stances philosophers and linguists have taken to a series of sentences. Consider first:

(1) `Cicero admired Tully.`

(Throughout this discussion, I'll treat `admire` and other transitive verbs as extensional.) Consider next:

(2a) `Cicero admired him` (pointing at a bust of Marcus Tullius Cicero, here also claimed to be doing the admiring).

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

or:

(2b) He (pointing at the bust) admired Cicero.

Sentences (2a) and (2b) do not have to be produced, or understood, in ignorance of the fact that Cicero is the one being demonstrated. They might be embedded in a larger discourse like so:

(2c) Cicero was too proud to admire other people, but Cicero admired him (pointing), so he (continuing to point) must be Cicero.[1]

Contrast the way the pronoun him works in (2a) and (2c) with the way it works in:

(3a) Cicero divorced the woman who bore him children.

Here him is meant to be, and should be understood to be, anaphoric on the occurrence of Cicero, and hence to *derive its value from* that antecedent expression.[2] If we generate a sentence that parallels (1) and (2a), but with this anaphoric relation between the pronoun and Cicero, we get:

(3b) Cicero admired himself.

There are complex grammatical constraints on when anaphoric pronouns must or must not take reflexive morphology (himself rather than him). For our purposes, observe that pronouns can be understood as anaphoric without the presence of that morphology, as in (3a), or:

(3c) Cicero's wife disappointed him.
(3d) Terentia betrayed Cicero, so he was unhappy.
(3e) Cicero mourned his daughter.

Examples like these are sometimes hypothesized to be ambiguous. For example, (3e) might state that Cicero has the reflexive property $\lambda x.\ x$ mourned x's daughter, or that he has the relational property $\lambda x.\ x$ mourned Cicero's daughter. The latter reading would be required if the his were demonstrative and just happened to designate Cicero, as in (2a) and (2b); but it's also available when his is anaphoric. If there are these two readings of (3e), they'd underwrite an ambiguity that's widely agreed to be present when the sentence continues:

(4a) Cicero mourned his daughter, but Atticus didn't.

Are we talking about Atticus mourning Atticus's own daughter, or about his mourning the same person Cicero did? For historical reasons, linguists call the first reading "sloppy" and the second "strict."[3] The "sloppy" reflexive *reading* of these sentences should not be confused with the reflexive *morphology* exhibited in (3b). As we've just seen, the reading can be present without that morphology. Further, the morphology doesn't force that reading:

(4b) `Cicero admired himself, but Atticus didn't.`

can also be read "strictly," as saying that Atticus didn't admire Cicero.[4]

Our stage consists of stances philosophers and linguists have taken to examples like (1) and (2ab), on the one hand, versus examples with anaphora like (3a–e), and perhaps also:

(5) `Cicero admired Cicero` (meant and understood to involve a recurring use of a single name).

Bracketing examples like (5) for a moment, one stance toward the (3) examples is represented by Lasnik (1976) and Bach (1987, chs 11–12). These theorists deny that anaphoric relations generally have any distinctive syntactic or semantic manifestation. In essence, they'd treat the (3) examples on the same model as (2ab).[5]

A second stance is exemplified by Salmon, Soames, and some other theorists. They embrace three commitments, the first of which I'll label:

(Z1) So far as possible, analyze "sloppy" reflexive readings using bound variables.

That is, they'd take (3b), when so read, to have the form:

(3b′) ($\lambda x.$ *x* admired *x*) Cicero

It's controversial how far such analyses can be applied. Salmon and Soames deny that (5) has this form. That denial is consonant with the fact that continuing (5) with `but Atticus didn't` only allows a "strict" reading.

Restricting our attention to examples that *do* have "sloppy" readings, it's controversial whether a bound-variable analysis can account for all of them. (3d) is less hospitable to such an analysis than (3abe), and examples with "donkey pronouns" like:

(6) `Every orator who had a daughter loved her.`

are widely agreed to resist it: `her` is outside the syntactic scope of any binding introduced by `a daughter`. Yet such sentences still support "sloppy" readings:

(7) `Every orator who had a daughter loved her, but some consul who had a wife didn't.`

meaning the consul did not love his wife.[6]

Salmon and Soames also hold:

(Z2) Propositions of the form ($\lambda x.$ *x* admired *x*) *a* are distinct from propositions of the form ($\lambda x.$ *x* admired *a*) *a* and ($\lambda x.$ *a* admired *x*) *a*.

and:

(Z3) Attitude verbs like `believe` express a dyadic relation between a subject and the proposition expressed by their complement clause.

In light of (Z2), this has the consequence that attitude reports like (8ab) differ in meaning (and, on their view, in truth-conditions) from reports like (9abc):

(8a) `Anita believes that someone admired himself.`
(8b) `Anita believes that (λ`*`x`*`. `*`x`*` admired `*`x`*`) Cicero.`
(9a) `Someone is such that Anita believes that he admired him.`
(9b) `(λ`*`x`*`. Anita believes that `*`x`*` admired `*`x`*`) Cicero.`
(9c) `Anita believes that (λ`*`x`*`. `*`x`*` admired Cicero) Cicero.`

When the binding quantifier or λ operator is inside the complement, as in (8ab), Salmon and Soames say the report relates Anita to a content from which (in any guise) she could infer:

(10) `Someone admired himself.`

When it is outside the complement, as in (9ab), or the binding pattern is as in (9c), they deny that such a content is attributed to Anita. For all that (9abc) say, on their view, Anita may merely be willing to assent to sentences like (1).[7]

At several points in this discussion, I will need to invoke this notion of "contents from which one could infer (something like) (10)." Let's call them "cyclic" contents and thoughts.[8]

The third stance in our drama takes inspiration from an idea of Putnam's that (1) and (5) have different "logical structures." This idea was cultivated by Mark Richard in the 1980s and 1990s, and has been developed in different shapes more recently by Kit Fine and by others.

This stance agrees with Salmon and Soames that there's interesting semantic behavior present in (at least some readings of) examples like (3b) that's absent from (1) or (2ab). But these new theorists recognize among such behavior a pattern that extends also to the "strict" readings of the anaphoric examples, and may also be present in (6) and/or (5). Roughly, their idea is that *all* these examples have cyclic contents, whereas with (1) and (2ab) no inference to a claim like (10) is semantically underwritten.

The sense in which claims like (10) may be "inferable" for a subject, or "semantically underwritten," is not straightforward. This must mean something more than a mere metaphysical entailment, since we already have that with (1). It must also mean more than that the consequence is intended by the speaker, or follows given what's presupposed in the discussion (see note 1 above). With (2c) we'd already have (10) be entailed in those senses. Instead, claims like (10) presumably must follow *from the conventional meaning* of the sentences we're considering (when used in the way we're considering), in a way that full competence with the sentences (so understood) requires one somehow to be sensitive to. OK, but what does that "requirement to be sensitive" amount to? At this juncture, proponents of the third stance reach for ideas like "what a speaker can know *a priori*," or "what a speaker would be immediately justified in believing" on the basis of their understanding. But the first idea threatens to let in all sorts of substantive mathematics that shouldn't be

part of the linguistic competences we're trying to explicate. And the second idea provokes worries about speakers who have no semantic concepts, or speakers who do but endorse false semantic theories (perhaps justifiably).

We will return to issues about the epistemology of language use later. Rather than begin with any specific theory here, I propose we just content ourselves for now with an intuitive understanding of the proposal that the (3) examples are cyclic on both their "sloppy" and "strict" readings, that is, that they *somehow* semantically underwrite inferences to claims like (10). Proponents of this stance can allow "sloppy" and "strict" readings to have different contents, but they'll say that the behavior just described is at least interestingly shared by them; and they'll insist that the "strict" readings have contents different from those in the (2) examples. As I said, this same strategy may be extended to other examples like (6) and/or (5).[9]

The syntax and semantics of our (3) examples is contested, so it's not obvious that the "strict" readings of these examples *need to* end up with their own distinguished contents, from which claims like (10) could be inferred. But that is the view our third stance will defend.

If patterns like the ones we're considering are acknowledged to be semantically encoded, what difference should that make, for example to the truth of attitude reports that embed these sentences as complements? This is a contested issue that we'll explore below.

Here are some more examples that may exhibit the envisaged patterns, where this also can't be the result of the kind of binding structure exhibited in (3b′). First:

(11a) `Only Cicero mourned his daughter.`

This is ambiguous between:

(11b) `Cicero, and no one else, mourned the daughter of Cicero.`
(11b′) (λ*x*. *x* mourned Cicero's daughter) Cicero; and no one else _____

and:

(11c) `Cicero, and no one else, mourned that person's own daughter.`
(11c′) (λ*x*. *x* mourned *x*'s daughter) Cicero; and no one else _____

where the _____ indicates that a "matching" predicate has been deleted or elided, namely the λ-term from the first conjunct. Yet it's not only (11c) that conventionally encodes that it's the same person who both mourned and was the father. The other reading should also encode this.[10]

Second:

(12) `Terentia didn't admire Cicero, but he/Cicero did.`

It's most natural to analyze this as:

(12′) ¬(λ*x*. *x* admired Cicero) Terentia; but _____ Cicero

Not as:

(12″) ¬(λ*x*. *x* admired Cicero) Terentia; but (λ*x*. *x* admired *x*) Cicero

because on current theories of ellipsis, such non-matching λ-terms would not license the surface form (12).[11] Nor should we analyze it as:

(12‴) (λ*y*. ¬(λ*x*. *x* admired *y*) Terentia; but _____ *y*) Cicero

That would posit a syntax and predicate structure more complex than we have independent evidence for. Yet the second conjunct of (12) does seem to conventionally encode that it's the same person who both admires and is admired. So it should be possible for this to happen even when the binding structure is as in (12′).[12]

Third:

(13) `Terentia preferred Cicero to every politician that he/ Cicero did.`

It's most natural to analyze this as:

(13′) ∀*p* : _____ Cicero. (λ*x*. *x* preferred Cicero to *p*) Terentia

Not as:

(13″) ∀*p* : (λ*x*. *x* preferred *x* to *p*) Cicero. (λ*x*. *x* preferred Cicero to *p*) Terentia

Yet the restriction on the quantifier of (13) does seem to conventionally encode that it's the same person who both prefers and is ranked first. So it should be possible for this to happen even when the binding structure is as in (13′).

In summary, the third stance thinks we need to go beyond the bound-variable analysis to capture the pattern they're attending to, and want to argue is semantically encoded. This chapter will survey some further motivations for, and distinctive claims made by, this stance.

2

Let's go back to the early 1950s. Carnap had recently proposed that sentences are synonymous, and express the same belief, iff they are "intensionally isomorphic": that is, they have parallel constructions with corresponding atomic elements always having the same intension. Carnap had also proposed that a report in English attributing `believes that` *S* is true iff there is a language *L* and sentence *T* such that *S* in English is intensionally isomorphic to *T* in *L*, and the subject is disposed to assent to *T* as a sentence of *L*.

In the early 1950s, Church and Mates pressed several objections against these proposals. Among other things, Church (1950) complained that Carnap's analysis wrongly characterizes ordinary beliefs as being "about sentences"; he preferred analyses that instead relate

believers to propositions. Such analyses would translate more convincingly into other languages. Church (1954) argued that intensional isomorphism is anyway too weak for Carnap's purposes, since it tolerates the substitution of non-synonymous expressions with the same intension. Yet such substitutions seem to change the truth-value of attitude reports. Church favored working with a notion of "synonymous isomorphism" instead. Mates (1950) complained that even with a tighter synonymy relation between distinct sentences *S* and *S′*, one can still always embed them:

(14a) `Whoever believes that S, believes that S.`
(14b) `Whoever believes that S, believes that S′.`

in such a way that speakers can intelligibly assent to one embedding while doubting or dissenting from the other. Here is an example:

(15a) `Whoever believes that lawyers are wealthy, believes that lawyers are wealthy.`
(15b) `Whoever believes that lawyers are wealthy, believes that attorneys are wealthy.`

I know that lawyers and attorneys are the same. But conceivably, some subjects may suspect that they aren't. I may regard such subjects as counter-examples to (15b), and for that reason be unwilling to assent to it. Carnap's strategy (even as refined by Church) must count subjects like them (and me) as nonetheless *having* beliefs reportable with complement clauses that they'd resist assenting to.[13]

Putnam (1954) defends Carnap's strategy against some of these objections. Against Church, he claims that multiple analyses can be correct and yet not intertranslatable. Crucial for our purposes is Putnam's response to Mates, which involves re-defining "intensionally isomorphic" so as to be sensitive to differences like those exhibited between (15a) and (15b), or (5) and (1). Here is Putnam's own example:

> "Greek" and "Hellene" are synonymous. But "All Greeks are Greeks" and "All Greeks are Hellenes" do not *feel* quite like synonyms. But what has changed? Did we not obtain the second sentence from the first by "putting equals for equals"? The answer is that the *logical structure* has changed. The first sentence has the form "All F are F," while the second sentence has the form "All F are G" – and these are wholly distinct … (Putnam, 1954, pp. 153–154)

Putnam's idea here is that the meaning of a sentence is not *just* a function of the meaning of its parts, but also of the sentence's "logical structure," and that Mates had been assuming too simple an account of that. If we instead work with a notion of structure that's sensitive to the recurrences of terms (see Putnam's n. 10), so that the difference exhibited by `All F are F` and `All F are G`, and presumably also by `Cicero admired Cicero` and `Cicero admired Tully`, are semantically significant, then these sentences as wholes will be non-synonymous, even though their individual words pairwise are synonymous.

This idea of Putnam's was seconded by Kaplan and elements of it can also be found in Geach.[14] Below, I'll first explain the modern development of the idea in Richard's and Fine's work, and then §5 will more briefly survey some other places it appears. The authors we'll consider don't all work out Putnam's guiding idea in the same way, and they use varying terminology: "logical potential" (Taschek), "co-relativized" terms (Richard), "coordination,"

"representing as the same," and "strict coreference" (Fine), "explicit coreference" (Taylor), "coindexed expressions" and "grammatically determined coreference" (Fiengo and May), and "*de jure* coreference" (Schroeter, Pinillos, Recanati). The last phrase seems to be trending most strongly, so I've based this chapter's title on it. I prefer to say "codesignating" rather than "coreferring," though, as to some the term "referring" has specific connotations that aren't essential to the phenomena we're exploring (even after we restrict our attention to singular terms, as we will here). These readers would resist talk about variables or anaphoric pronouns *referring*. "Designating" is more readily understood to have the needed generality.

3

Richard's engagement with Putnam's idea begins in his (1983). This paper became famous for its phone-booth scenario, which presented an intuitive difference in acceptability between:

(16a) `I believe that you are in danger.`
(16b) `The man watching you believes that you are in danger.`

despite the fact that these differ only in the substitution of codesignating expressions *outside of* any opaque context.[15] That point was only subsidiary to Richard's main goal in the paper, though, which was to explore the idea that:

(17a) `I can inform you of her danger via the telephone.`
(17b) `I can inform her of her danger via the telephone.`

differ semantically in such a way as to not be substitutable in the complements of attitude reports. For Richard at this time, sentences (17a) and (17b) do have the same *content*, but he makes the attitude verbs they may be embedded under become sensitive to their differing *character*. (We won't dwell on the details of how this works, because as we'll see, Richard soon moved to a different and more general strategy.)

Sentences (17a) and (17b) are of the form `...you...her...` and `...her...her...`, where all the pronouns are codesignating, and thus they parallel our examples:

(1) `Cicero admired Tully.`
(5) `Cicero admired Cicero.`

The only differences are that in Richard's examples, the pronouns occur in what may arguably be an intensional context (`inform _____ of her danger`) – this plays no role in his discussion – and his examples concern pronouns rather than proper names. This limitation was essential to the proposal he advanced in the 1983 paper, but was lifted when he developed these ideas further (and differently) in his 1990 book.

Richard (1987) begins to extend his claims about (17ab) to the open (quasi-English) expressions:

(18a) `Juan said that (he observed` *x* `and then` *y* `but he wanted to observe` *y* `and then` *x*`).`
(18b) `Juan said that (he observed` *x* `and then` *y* `but he wanted to observe` *x* `and then` *y*`).`

and:

(19a) Juan said that (*x* chased *y* and then *y* chased me).
(19b) Juan said that (*x* chased *y* and then *x* chased me).

arguing that each pair may differ in truth-value even when the variables *x* and *y* are assigned the same designation. The formal mechanisms Richard proposed in (1983) won't explain this result, since the variables don't differ in character. We don't get Richard's developed account of how this might work until Richard (1990). (Similar examples are discussed there at pp. 153, 200ff.)

Richard (1986–1987) offers an intermediate story, introducing a construction he calls "R-structures" (these are akin to the DRSs used in Kamp, 1984–1985; 1990; and Asher, 1986; 1987; 1989). We won't dwell on the details of these either. But I do want to point out several elements in this paper that will reappear in later contributions by Richard and others. First, Richard proposes:

> Let us talk of two occurrences of a term being co-relativized, when they are treated as if they were occurrences of the same existentially bound variable.
>
> (Richard, 1986–1987, p. 251)

Second, Richard associates several "levels" of information with a report (1986–1987, pp. 253ff.). The first level is initially described as just the Russellian content of the complement (pp. 244, 253) but then later is said to be that complement's R-structure (p. 256). This means that it will encode patterns of how terms recur in that complement. The second level encodes patterns of how terms recur in *multiple* reports, within a larger discourse. This roughly corresponds to Fine's "weak *de dicto*" reading of reports, which we'll discuss later. The third level encodes patterns of how terms recur *also in our unspoken background convictions* about the attitudes of others (pp. 257ff.). We will see something like this in Fine and other authors, too. In this paper, Richard only regards the first of these levels as giving the strict truth-conditions of the report (pp. 244, 256).

The account of attitude reports Richard arrived at in his (1990) is complex, and will be best understood against the background of other views in the air at the time.

Instead of the Fregean strategy of having beliefs and other attitudes be relations to a single fine-grained object, many philosophers in the 1980s began following Kaplan's and Perry's suggestion to think of attitudes as relations to *both* a coarse-grained Russellian proposition, *and* something that added additional grain, such as a *guise* under which that proposition is grasped, or a *sentence in Mentalese* which expresses it, or a series of *mental files* deployed in thinking it.[16] For concreteness, let's work with the Mentalese sentence version of this idea. So the underlying facts about a subject, that a belief report would aim in some way to describe or summarize, would be that he stands in some belief-like accepting/endorsing relation to a coarse-grained Russellian proposition about Cicero and admiring *with* some Mentalese sentence like `Cicero admires Tully`. *Exactly how* belief reports aim to describe or summarize these underlying facts, different theorists gave different accounts of. To help separate what these accounts agreed about from the different proposals they offered of the predicate `believe`, let's call the underlying belief-like accepting/endorsing relation "doxizing." One doxizes a Russellian proposition with a Mentalese sentence.

Salmon (1986a) offered a straightforward account of the relation between doxizing and the predicate `believe`: in the terms we're employing here, his view was that a report:

(20) `Anita believes that` *S*.

is true iff *there exists some* Mentalese sentence *T* such that Anita doxizes the Russellian proposition expressed by *S* with that sentence *T*.[17] Salmon is not always explicit whether this is meant to be an account of the *logical form* of the belief report, or only its truth-conditions. On the first construal, belief reports would themselves contain existential quantifiers over Mentalese sentences (or guises); these would just have no overt pronunciation. Also, the predicate `believe` would itself really contribute a triadic semantic value, contrary to its surface appearance. Some have interpreted Salmon this way;[18] but the balance of evidence speaks more strongly for the other construal, where `believe` really only contributes an atomic dyadic relation, as it appears to (see Salmon, 1986a, pp. 5–6, and clauses 35–36 on pp. 146–147; also his 1986b, pp. 32–33), and the existential quantifier only enters into Salmon's explanation of the report's *truth-conditions*, not its actual syntax or semantic structure.

Many theorists agreed with Salmon that in at least some cases, belief reports are completely non-committal about how (with what kind of Mentalese sentence) their subject doxizes the attributed proposition. But these theorists resisted saying that it was *always* so. In many other cases, they claimed, belief reports make more specific commitments about the third argument of the doxizing relation. Many went no further than that, saying only that the particular complement clause used in the belief report is meant to somehow "display" (without designating or literally expressing) the kind of Mentalese sentence the subject doxizes with.[19] But a handful of theorists made specific proposals about how belief reports constrain or supply the extra argument.

The earliest of these proposals came from Schiffer, who suggested that (20) be analyzed as:

(20′) $\exists T\,(\phi T \wedge \text{Anita doxizes } \sigma \text{ with } T)$

where σ is the Russellian proposition expressed by *S*, and *T* is a Mentalese sentence or guise meeting some condition ϕ (see esp. Schiffer, 1977; 1992). Because ϕ is a contextually supplied element that isn't overtly designated, views of this style became known as "hidden indexical theories" (HITs). I will refer to this as "Schiffer's HIT," though he doesn't straightforwardly endorse the view. He merely says it's the best option if you want to embrace compositionality about what complement clauses express. Over the course of many papers, Schiffer voices a series of objections against "his" HIT.[20] Three of these I'll mention just in passing: we shouldn't advocate such a view for speech reports, there are no good candidates to play the role of "guises," and the view would need supplementation to deal with empty names. Three other objections it would be useful to refer to later, so I'll label them:

S1 the "meaning-intention" problem: Ordinary speakers aren't aware of quantifying over, describing, or designating guises. So it's implausible that they have the semantic intentions they'd need to have, if a HIT of attitude reports were correct.

S2 the "logical form" problem: Schiffer's HIT analyzes `believe` in terms of the triadic predicate `doxize`. But Schiffer says that syntactic evidence, and ordinary intuition, speak against analyzing `believe` as having more arguments than just a subject and a complement (plus perhaps a time, which we're here ignoring). This complaint is discussed further in Ludlow (1995), Schiffer (1996), and Ludlow (1996).

S3 the "validities" problem: The argument `Anita believes everything Juan says; Juan says that Cicero is alive; so Anita believes that Cicero is alive` seems intuitively to be valid. But if the first premise says merely that if Juan says σ, Anita will doxize σ under *some guise or other*, then the premises don't entail a conclusion that constrains how Anita doxizes. If on the other hand, the first premise says that Anita will doxize σ *under the same guise that Juan says it*, then the conclusion follows only if it involves the same guise that's invoked in the second premise. And Schiffer's HIT doesn't guarantee that it will. (On his account, the premise and the conclusion will each have their own $\exists$.)

Crimmins and Perry *do* endorse a version of HIT.[21] On their account, though, speakers often directly specify the subject's Mentalese sentence *T*,[22] rather than contextually supplying a condition ϕ that restricts a quantifier over *T*. Though *T* is directly specified in *that* sense, it still is not designated by any pronounced syntax: in Perry's terminology, it's an "unarticulated constituent" of the report.[23]

Recanati (1993, chs 18–20; 1995) defended a view with important similarities to Schiffer's and Crimmins's HITs, but also some notable differences (Crimmins, 1995c, pp. 201–203, compares their views). Like the others, Recanati wanted belief reports to often assert, and not merely implicate, constraints on how (with what Mentalese sentence or guise) the subject doxized. Also like the others, and unlike Fregeans, he wanted component expressions of the report's complement to have their ordinary designations – thus making it unproblematic to say things like `Anita believes that Cicero is alive, but he isn't`. The strategy Recanati pursues to reconcile these goals is not to make the predicate `believes` take a third argument, which the report quantifies over or contextually supplies. Instead, he proposes that the complement clause as a whole usually expresses an enriched, "quasi-singular" proposition that correlates ordinary designata (like Cicero) with the subject's Mentalese representations of them.[24] For example:

(21) `Anita believes that Cicero is alive.`

might express:

(21′) ⟨a dyadic believing relation, Anita, ⟨the property of being alive, (Cicero, *the mental representation of Cicero that Anita associates with the name* `Cicero`)⟩⟩

Recanati denies that the enrichments are determined by the conventional meaning or the truth-conditional content that `Cicero is alive` would express in isolation. Rather it depends on the context of the report, including previous conversational moves and other specifics about how that sentence is embedded. (For example, it may depend on how we designate Anita.)

Recanati's view will help us contextualize Richard (1990). The spirit of the theory Richard develops there and in later work[25] is reminiscent of Carnap's account from §2 above:

> [W]hen we ascribe a belief saying `so and so believes that` *S*, we offer the sentence *S* as a *representation* or *translation* of what realizes one of so and so's beliefs. Attitude ascription thus presupposes something like a "translation manual," one keyed specifically to the individuals to whom attitudes are ascribed … The semantic rule governing belief ascriptions is something like: *x* `believes that` *S*, used in a context *c*, is true just in case *S*, relative to context *c*'s translation manual, translates some belief-realizing state of (the referent of) *x*.[26]

Concerning Richard's use of "translate," this should not be understood to mean that the expressions in the report have to be *conventional* translations of any expressions the subject would accept (see, e.g., 1990, pp. 134–135). Concerning Richard's use of "represent," this can also be confusing. He uses that term in several ways. "Representations" are groupings of token Mentalese expressions from which (what I'll call) private RAMs are composed (pp. 181–190; this is defined below). This is related to Richard's "representational systems," which are sets of all a subject's private RAMs (p. 137). Another use of "represent" (invoked in the above quote) is for the relation that a public RAM stands in to the private RAM that is its image under a given correlation function (pp. 139, 144). Importantly, Richard (unlike Crimmins) doesn't think that reports *designate* or directly specify subjects' Mentalese representations.

We need to look at the details of how Richard spells this out. They place him somewhere between Schiffer and Crimmins, on the one hand, and Recanati, on the other. Like Recanati, Richard wants complement clauses to express something richer than mere Russellian propositions. But unlike Recanati, he wants a compositional story about what the complement clauses do express. And unlike Recanati and Crimmins, he thinks reporters generally don't transact with other subjects' specific Mentalese representations. So rather than have the complement of (21) express a proposition enriched with Anita's Mentalese, Richard proposes it's instead enriched with *the English name* `Cicero` that was used in the report.[27] At first (1989; 1990; 1993) he called these enriched propositions "Russellian Annotated Matrices" (RAMs), but later he calls them "articulated" (1995) or "sentential" propositions (2006). He calls the components of these RAMs – the pairings of individual words with their designations – "annotations."

How do we get from a proposition enriched with English vocabulary, which the report's subject need not understand, to information about how the subject does mentally represent Cicero? Let's call the RAM expressed by the report's complement the "public" RAM. Let's call the Mentalese RAMs, with which subjects are claimed to doxize, "private" RAMs. Richard posits "correlation functions" that take us from the public RAM to a private RAM. We can understand him as proposing that (21) expresses:

(21″) $\exists f\,(\phi f \wedge \text{Anita doxizes}^{\star}\ f(\sigma^{\star}))$

Here, doxizing* is a dyadic relation Anita stands in to private RAMs that she "accepts."[28] $\sigma^{\star}$ is the public RAM expressed by the complement `Cicero is alive`: the fusion of a Russellian proposition with those very English words. f is a correlation that takes us from $\sigma^{\star}$ to some private RAM. The report doesn't directly specify f, but it does supply restrictions ϕ, in a context-dependent and not-overtly-pronounced way. In this regard, the view resembles Schiffer's HIT.

I've introduced Richard's theory this way to highlight respects in which it agrees with Schiffer, Crimmins, and Recanati. But in fact this is not quite the form that Richard himself employs. Rather than have the context supply an argument ϕ, Richard instead folds the ϕ-role into the relation I called "doxizing*," giving a single (contextually varying) relation that I'll call "doxphizing*." So Richard's own version of (21″) looks more like:

(21‴) $\exists f\,(\text{Anita doxphizes}^{\star}\ \sigma^{\star}\ \text{under}\ f)$

The truth-conditions of this are the same as (21″). Because doxphizing* includes the contextual restrictions on acceptable correlations, Richard's account is one where the predicate

`believe` expresses different relations in different contexts. (On Schiffer's and Crimmins's HIT, by contrast, `believe` is not context-sensitive, though the report as a whole is; and on Recanati's account, it's the report's complement that's context-sensitive.)

Let me point out some details and refinements.

- Richard understands his *correlations* as functions that map annotations to annotations, in a way that in general must preserve Russellian content and sometimes has to also obey other contextual restrictions (captured above by ϕ or doxphizing*). What will be important for our discussion is that these are *functions* part of whose effect is to map *expression types* in the report language to types of Mentalese expressions (see Richard, 1990, ch. 3, n. 11; Crimmins, 1992a, pp. 191–192; Sider, 1995, p. 503; Soames, 1995, n. 12; 2002, ch. 7, n. 12).

  A consequence of this is that when (non-demonstrative) expressions recur in the report's complement, as `Cicero` does in:

  (5*) `Anita believes that Cicero admired Cicero.`

  then the report commits to the subject also having recurring Mentalese expressions in her private RAM (see Richard, 1990, pp. 138–140, 201–202, 217–219). Hence, (5*) attributes a cyclic thought to Anita, one from whose content she could infer:

  (10) `Someone admired himself.`

  This prediction should seem familiar from our summary of Richard's three earlier papers, at the start of this section. It corresponds to a claim I'll label F3 when discussing Fine. (Richard also says that complements containing pronouns like `himself` always attribute cyclic thoughts: 1990, p. 218.)

  As we'll discuss, these commitments are controversial. For example, we observed that Salmon and Soames only allow complements like (3b)/(3b′) to attribute cyclic thoughts. They deny that the complements of (9bc) or (5*) have that structure.

  The rest of Richard's account does not *force him* to understand correlations in the way described here. He could instead define them on expression *occurrences* (see Crimmins, 1992a, pp. 195–196; 1995a, p. 387), or he could work with *relations* here instead of functions (see Sider, 1995, §6). Then recurring expressions in a complement needn't attribute cyclic thoughts. But Richard made this choice deliberately: in part to underwrite the truth-conditions he predicted in the earlier papers, and for further reasons he spells out in (1993, pp. 117ff.).

  Eventually, though, Richard came to have second thoughts about this, and in (2013b, pp. 8–11) he suggests that for iterated reports, the commitment described above only holds for *one reading*. He also countenances other readings, that permit recurring expressions in the complement to be "indexed to a subject higher in the sentence," and thus to possibly "translate" diverse Mentalese expressions from a more proximal subject's private RAM.[29]

- The fact that Richard's presentation makes `believe` indexical gives him other troubles with iterated reports. Suppose I observe Juan asserting (21), and then, in another context, report `Juan said that Anita believes that Cicero is alive`. Since my context is not the same as Juan's, it's unlikely that we have the same restrictions on acceptable correlations. That means that `believe` in my mouth won't express the same doxphizing* relation that it did in Juan's mouth. This motivates Richard to propose

that `believe` in the complement of my report contributes *its character* rather than its content to my report's public RAM (1990, pp. 165–167, 245–246; 2006, pp. 259–60). (And so too for other attitude verbs.) This is why I said above that correlations "*in general* must preserve Russellian content": here we have some exceptions, where it's instead the expression's character that needs to be preserved.

This raises several issues. First, why make a special exception just for attitude verbs? Richard ends up suggesting a parallel treatment of gradable adjectives and some uses of expressions like `foreign` and `the local` (2006, pp. 260–262; 1993, n. 3). Second, this aspect of Richard's account leads him to a *relativist* view of attitude reports (1990, §4.4), where *the content* of a report needs a further contextual parameter before it is truth-evaluable. Richard welcomes this commitment (see further his 2008; 2015, chs 5–7), but others may not.

I will ignore this aspect of Richard's theory.

- In (21″) and (21‴) we had an $\exists f$ in the content of a single report. In fact Richard's considered view is that sometimes we only quantify over correlations *at the discourse level*, and thus might reuse the same correlation f in several reports (1990, pp. 143, 175–180, 235–243; see also 1986–1987, pp. 69ff.). This corresponds to a move I'll label F4 when discussing Fine, and helps answer Schiffer's objection S3 to HITs, mentioned above (see Richard, 1986–1987, pp. 69, 75; 1990, pp. 148–149; 1993, pp. 118–120; contrast Crimmins, 1995a, pp. 392ff. esp. n. 11). It also means that Richard is proposing that `believe` expresses a *triadic* relation (with an extra argument place for the correlation f supplied by the discourse). So he has to confront Schiffer's objection S2.

Most of the critical attention to Richard's theory focused on difficulties about how context is supposed to supply restrictions on correlations. Some authors made complaints reminiscent of Schiffer's objection S1 to HITs (see Saul, 1993; Clapp, 2000; Soames, 2002, ch. 7; also Richard, 1997, §11). Soames (1995; 2002, ch. 7) complained that Richard's initial account of restrictions gave reports implausible modal profiles. Sider (1995, §§2–4) and Soames (1995; 2002, pp. 179–191) described cases where conversants unwittingly accept conflicting restrictions in a single context, because they don't recognize a subject whose attitudes they're reporting under different guises (see also Nelson, 2005). These complaints pose serious challenges to Richard's theory, but they are orthogonal to the aspects of his theory that are interesting for this chapter, so I won't pursue them.

A handful of other critical discussions merit special mention. First is Soames (1987b, esp. §§6, 8). This addresses an early undeveloped version of Richard's views, but many of the complaints raised there are also ones Soames will raise against *Fine's* Putnam-inspired account. Second is Soames (1994), which discusses "sloppy" donkey anaphora (pp. 120ff.) and some examples we'll return to in §6 below. The third critical focus is a complaint that Church (1954, p. 165) made against Putnam's original proposal. Given the different "logical structures" that Putnam attributes to `All Greeks are Greeks` and `All Greeks are Hellenes`, the latter would turn out to be *inexpressible* in a language that has only a single predicate synonymous with `Greek` and `Hellene`. Church found this intolerable. I'm not sure that judgment is underwritten by our actual practice and assessment of translation. But it is a judgment that Salmon shares and attaches great weight to (Salmon, 1986a, ch. 4, n. 4; 2001, pp. 582ff.; 2010, n. 11; 2012, pp. 437–438). See also Soames (1987b, pp. 113–114). Richard addresses those kinds of complaints at (1990, pp. 155, 160–162, 167–171).

4

(Unless otherwise indicated, page references in this section are all to Fine 2007.)

Fine's engagement with Putnam's idea[30] began in his (2003), which developed into chapter 1 of his 2007 book. The problem that concerns him there is how to account for the semantic role of the variables *x* and *y*. One the one hand, they seem to function semantically the same. We understand the two sentences:

$$\forall x Fx \quad \forall x Fy$$

to have the same meaning, and so too the two predicates:

$$\lambda x.Fx \quad \lambda y.Fy$$

and arguably we should also so regard the two open formulas:

$$Fx \quad Fy$$

and the mere variables themselves:

$$x \quad y$$

On the other hand, the predicates:

$$\lambda x.Rxx \quad \lambda y.Rxy$$

have different meanings, and arguably so too should the open formulas:

$$Rxx \quad Rxy$$

Fine's way of distilling this issue is to say that whereas the two single variables *x* and *y* have the same semantic role, the *pair* of variables *x*, *x* has a *different* semantic role from the pair *x*, *y*. Fine explores existing accounts of the semantics of variables and argues that none of them do justice to that idea in a philosophically satisfying way. For example, one account will give each of the variables *x* and *y* as their semantic value the set of objects they range over. But applying that to the pairs *x*, *x* and *x*, *y* would give them the same meaning. We could solve that by giving *x* as its meaning instead *the pair of itself plus* the set of objects it ranges over; and similarly for *y*. Or less artificially, we could give *x* as its meaning a function from assignment functions *g* to the object that *g* assigns to *x*; and similarly for *y*. Those approaches *would* give the pairs *x*, *x* and *x*, *y* different meanings, because they'd give each of the variables on its own a different (though analogous) meaning. Fine's objection to this is not a *direct* insistence that *x* and *y* on their own must get *exactly the same* meaning. Rather it's that such proposals are philosophically unsatisfying, because:

> [T]he posited difference between the semantic values for *x* and *y* simply turns on the difference between the variables *x* and *y* themselves … [W]hat we secure on this approach, strictly speaking, is not a *semantic* difference, one lying on the non-conventional side of language, but a *typographic* difference, one lying purely on the conventional side of language. (p. 11)

Fine criticizes other proposals in a variety of ways.

His own solution gives up the idea that expressions can be given any meaning *on their own* that settles what the meaning of pairs (or longer sequences) of expressions is. Instead of determining the meaning of *x*, *y* as built from the intrinsic meanings of its parts, he proposes a theory that determines the meaning of *x*, *y* as in the first place given by how those expressions are *semantically related* (see esp. pp. 3, 22–24). Note that a relation between *x* and *y* (such as codesignating Cicero) can be re-factored into a function from the pair *x*, *y* to the presence or absence of the relevant semantic attribute; thus we can also understand Fine to be assigning semantic values to *sequences* of expressions, rather than (only) to expressions taken individually. (Fine prefers to call these "semantic connections" rather than "semantic values," but I will stick with the more traditional vocabulary.) As Fine's account evolves to handle binding variables with quantifiers, we get the revision that it's not mere sequences of expressions, unadorned, that get assigned values, but rather these together with a "coordination scheme" mandating that some free occurrences of the same variables in the sequence be interpreted the same. (This is to capture the different ways that our interpretation of $(\forall xFx) \wedge Gx$ and $\forall x(Fx \wedge Gx)$ will depend on the interpretation of the sequence *x*, *Fx*, *Gx*.) Fine proposes to generalize the approach that assigns to each variable the set of objects it ranges over. Instead he will assign to each *coordinated sequence of* expressions a set of *sequences of* semantic values.

Fine does not give a full and explicit semantics for his language, but here is what I understand him to be proposing.[31] See also Pickel and Rabern (forthcoming, §2).

- $[\![\ldots, R, \ldots]\!]$ = {(…, *R*'s extension, …) | for each interpretation of the rest of the original sequence of expressions }
- $[\![\ldots, \underline{x}, \ldots, \underline{x}, \ldots]\!] = \{(\ldots, d, \ldots, d, \ldots) \mid d$ any object in the domain *x* ranges over, for each interpretation of the rest of the sequence }
- $[\![\ldots, x, \ldots, x, \ldots]\!]$, where the occurrences of *x* are not linked = {(…, d_1, …, d_2, ….) | d_1 , d_2 any objects in the domain *x* ranges over, for each interpretation of the rest of the sequence }
- $[\![\ldots, Rt_1t_2, \ldots]\!] = \{(\ldots, \text{true}, \ldots) \mid [\![\Sigma]\!]$ contains (…, Δ, d_1, d_2, …) such that $(d_1, d_2) \in \Delta\} \cup \{(\ldots, \text{false}, \ldots) \mid [\![\Sigma]\!]$ contains (…, Δ, d_1, d_2, …) such that $(d_1, d_2) \notin \Delta\}$, where Σ is the coordinated sequence …, *R*, t_1 ,t_2, … (preserving any coordination links from the original sequence …, Rt_1t_2, … in the natural way)
- $[\![\ldots, \neg\phi, \ldots]\!]$ = {(…, true, …) | $[\![\Sigma]\!]$ contains (…, false, …)} ∪ {(…, false, …) | $[\![\Sigma]\!]$ contains (…, true, …)}, where Σ is the coordinated sequence …, ϕ, … (preserving any coordination links in the natural way)
- $[\![\ldots, \phi \wedge \psi, \ldots]\!] = \{(\ldots, \text{true}, \ldots) \mid [\![\Sigma]\!] \text{contains} (\ldots, \text{true}, \text{true}, \ldots)\} \cup \{(\ldots, \text{false}, \ldots) \mid [\![\Sigma]\!]$ contains one of (…, true, false, …), (…, false, true, …), or (…, false, false, …)}, where Σ is the coordinated sequence …, ϕ, ψ, . . . (preserving any coordination links in the natural way)
- $[\![\ldots, \forall x\phi, \ldots]\!] = \{(\ldots, \text{true}, \ldots) \mid [\![\Sigma]\!] \text{contains} (\ldots, d, \text{true}, \ldots)$ for every object *d* in the domain *x* ranges over} ∪ {(…, false, …) | $[\![\Sigma]\!]$ contains (…, *d*, false, …) for some object *d* in that domain}, where Σ is the coordinated sequence …, x, $\underline{\phi}$, … (preserving any coordination links from the original sequence …, $\forall x\phi$, … in the natural way, and additionally merging in further links between the new occurrence of *x* and any free occurrences of *x* in ϕ)[32]

Chapter 2 of Fine (2007) turns to giving semantics for languages with names or constants. There Fine aims to apply the fundamental ideas of the account just sketched to examples like our:

(1) `Cicero admired Tully.`
(5) `Cicero admired Cicero.`

and he also proposes a semantic difference between these, even if the expressions `Cicero` and `Tully` would have the same meaning taken individually.[33] An obstacle to this proposal is that whereas variables are associated with *a range* of values, and so we can "coordinate" different variable occurrences by mandating that their ranges vary in lock-step with each other, names take only *a single* semantic value. So we can't use the varying-in-lock-step mechanism to distinguish the semantic value of the pair `Cicero`, `Cicero` from that of the pair `Cicero`, `Tully`. Fine admits that this makes his use of "coordination" to talk about semantically required codesignation for arbitrary expressions a bit awkward (pp. 39ff.).

As a result, the semantics we get in Fine's chapter 2 looks different from the account sketched above. Fine here works with an extension of the familiar structured Russellian account of meaning.[34] The extension is that when an object appears several times in a structured proposition,[35] there is the option of adding a "coordination wire" linking those different appearances.[36] This would enable us to represent the meaning of (5) as an extended, possibly-wired structured proposition where the two appearances of Cicero *are* linked, and the meaning of (1) as a possibly-wired structured proposition where they aren't.[37]

Some comments about this.

- The presence of wires in the meaning semantically encodes that multiple arguments to a (possibly complex) predicate are supplied by the same value. In principle, one might also consider the possibility of semantically encoding that multiple arguments are supplied by *different* values. Soames mentions such an idea in (2012, p. 255) (but see also his n. 14; and Salmon, 2012, p. 409, which offer related but different ideas). This possibility is not pursued in any of the work surveyed here. In the possibly-wired structured propositions Fine is working with, the alternative to a semantic wire explicitly requiring several arguments to be the same is instead just semantic neutrality, not requiring the arguments to be the same but not forbidding it either.
- We need to distinguish the possibly-wired structured proposition where multiple appearances of Cicero are explicitly unlinked, from the coarser, Russellian proposition that *carries no information about* whether those appearances are linked.
- Fine's use of the term "uncoordinated" is not consistent. He most often uses this to mean (i) a possibly-wired structured proposition that explicitly lacks some link (pp. 54–55, 78, 83, 96, 111, and perhaps 69); but other times he uses it to mean (ii) a coarser, Russellian proposition that is silent about what arguments may be linked (pp. 52, 56–59, 77). Fine also calls (i) "negative coordination" (p. 56), and Salmon (2012, n. 40) calls it "withheld coordination."
- Fine doesn't immediately say how to integrate the treatment of names in chapter 2 with a semantics for quantifiers. King (2007) has an account of structured propositions using "wires" to capture which quantifiers bind which argument positions.[38] This is ostensibly

> a different use of wires than Fine is working with, but if Fine chose to say that multiple variable occurrences bound by a single quantifier should always be coordinated, then King's wires could be viewed as a special case of Fine's wires.[39]

When several expression occurrences have a meaning that positively links their designata in the way described here, Fine says that the expressions "strictly corefer," in the sense that it is *semantically required that* they codesignate (see pp. 43, 46–47, 50–51, 59, 123; and his 2010d); and he says also that they "represent their objects *as the same*," which he distinguishes from the mere attribution of identity to those objects (pp. 39–40). Generally, recurrences of a single name like `Cicero` will bear these relations, but that is not necessary nor sufficient for doing so.[40] Fine offers as a "good test" for the presence of these relations considerations like this:

> An object is represented as the same [by several expression occurrences] in a piece of discourse only if no one who understands the discourse can sensibly raise the question of whether it is the same.[41]

Fine doesn't officially introduce any operators or predicates that are truth-conditionally sensitive to the presence of semantic wires of the sort we've described. In fact, he says:

> [T]he coordinative aspect of the coordinated content of a sentence, such as "Cicero wrote about Cicero," is entirely lacking in any special descriptive or truth-conditional character and relates entirely to how its truth-conditions … are to be grasped. (p. 59; contrast ch. 4, nn. 10 and 11)

But one would naturally expect the semantic differences Fine posits between (1) and (5) to contribute in some way[42] to a difference in truth-conditions between attitude reports for which they are the complement clauses:

(1*) `Anita believes that Cicero admired Tully.`
(5*) `Anita believes that Cicero admired Cicero.`

And in chapter 4, Fine does discuss two readings of attitude reports on which (1*) and (5*) may have different truth-conditions. His "weak *de dicto*" reading has coordination among expressions in a report's complement – as in (5*) but not (1*) – iff cyclic thoughts are being attributed (pp. 91, 102ff.). His "strong *de dicto*" reading has an even tighter connection between the expressions in the complement and expressions in the subject's linguistic or mental repertoire. However, Fine acknowledges at several places that he hasn't given, and doesn't know how to give, a compositional semantics for these reports.[43]

One noteworthy feature of the example reports Fine discusses in his chapter 4 is that they sometimes involve coordination across the complements of multiple attitude verbs, perhaps even ones involving different subjects.[44] If we assume that all and only occurrences of the same variables are coordinated in the following, Fine would regard these reports as having different truth-conditions (on some readings):

(19c) `Juan said that` *x* `chased` *y*`, and Anita wondered if` *y* `chased me.`
(19d) `Juan said that` *x* `chased` *y*`, and Anita wondered if` *x* `chased me.`

even when *x* and *y* are assigned the same designation.

A full appreciation and assessment of Fine's views about attitude reports has to consider his account of how the analogue of linguistic coordination occurs in thought, too – what I'm calling cyclic thought-contents. Fine counts it as a strength of his approach that the same fundamental ideas drive the story of both linguistic content and cyclic thoughts. But for several reasons, including limited space, I propose to ignore this part of Fine's story, and to keep our focus on proposals about the semantics of public languages (whether natural or formal). Let's continue on with merely an intuitive grasp of cyclic thoughts. They are the kinds of thoughts that subjects who consider and assent to (5) normally manifest, but subjects in `Paderewski` cases don't. (I discuss these thoughts further in my 2016.)

I've mentioned several key elements in Fine's account that are referenced elsewhere in this chapter. Let's gather and label them:

F1 Officially, Fine doesn't introduce operators or predicates whose truth-conditions are sensitive to coordination in their arguments. But one assumes that a full story about his "weak" and "strong *de dicto*" attitude reports would involve such, presumably by way of a difference in what possibly-wired structured propositions are expressed by different complement clauses (though see note 42 above).

F2 There are natural reasons for Fine to say, as he seems disposed to say, that multiple variable occurrences bound by a single quantifier are always coordinated.

F3 On some (readings of) attitude reports, expressions in their complement are coordinated iff they attribute cyclic thoughts. I'll only be referring elsewhere to the left-to-right half of this claim.[45]

F4 Fine allows coordination across the complements of multiple attitude verbs (even ones with different subjects). Recall that Richard also aimed to achieve this, by saying we sometimes only quantify over correlations at the discourse level.

I've proposed already to ignore one central element to Fine's thinking, namely:

- his specific account of cyclic thoughts.

I will also pass over the connections Fine draws between semantic wires and semantic requirements (see the citations a page ago), as well as:

- how he connects these wires to ideas about what's "transparent" to a competent language-user, or his epistemology of language use more generally (pp. 49, 60–65). I'm sympathetic to many of Fine's claims about what's "semantically required," but would resist treating knowledge of such requirements as necessary conditions for competence or understanding.[46]

In a moment, I'll identify a third element in Fine's thinking that I'll also be ignoring. And in §7 below, I'll demonstrate a language that (*contra* point F1 above) does have predicates whose application is sensitive to semantic wires among their arguments.

Fine developed his account further in his (2010d), and in responses to his critics.[47]

Fine presents his account as a form of "Referentialism" or "Millianism" (pp. 5, 37, 45, 53), and contrasts it to Fregean accounts (pp. 35–37, 42, 58–60). Some of his commentators challenge this, and suggest his account may be closer to Fregeanism than it is to familiar

forms of Millianism (see also Richard's, 1990, pp. 147–154 comparison of his own view to Fregeanism). The taxonomic question here isn't especially interesting in itself.

More pressing is another complaint voiced by several critics, namely that Fine's account doesn't seem to offer any way to semantically distinguish:

(1) `Cicero admired Tully.`
(22) `Tully admired Cicero.`

or:

(23a) `Cicero was an orator.`
(23b) `Tully was an orator.`[48]

If these sentences occur as *part of a larger discourse containing other occurrences* of the names `Cicero` or `Tully`, then Fine's account as developed so far *does* distinguish discourses containing (1) from ones containing (22), and similarly (23a) from (23b) (see Fine pp. 52–53). But what if the sentences *aren't* part of such larger discourses (see Soames, 2010, p. 470; 2012, pp. 245, 250)? There is considerable pre-theoretic pressure to still see some semantic difference between (1) and (22), and reluctance to allow our story about how (1) relates to (22) to fundamentally differ from our story about how it relates to (5). Of course, standard Millians already deny there is any semantic difference between (1), (22), and (5). But they've built up a collection of explanatory tools to make those denials more palatable. Fine's account may have encouraged us to think we wouldn't need those tools; and if he after all needs to resort to them to accommodate intuitions about (1)/(22) and (23a/b), this may undermine some of the motivation he offered for the extra semantic machinery. This is a complaint Soames voices several times (in his 2010, pp. 473–474; 2012, pp. 234, 237–238, 262–263; see also his 1994, pp. 132–133).

One point Fine can make in reply is that his account would at least distinguish *embeddings* of (1) and (22) in attitude reports, on his "strong *de dicto*" reading of those reports. Admittedly, we're only offered a skeletal story about how that reading works. In Fine (2010b) he develops a more ambitious response, which involves coordination of expression occurrences in utterances of (1) not only to discourses in which it's *actually* contained, but also to all discourses in which it *might possibly* be contained.[49] If one gets over the extravagant multiplying of meanings it requires, this move offers powerful armament to Fine: it enables him to treat isolated utterances of (1)/(22) or (23a/b) in the same way he's already prepared to treat discourses containing them and other occurrences of `Cicero` or `Tully`. I worry if the armament may be *too* powerful, and no longer sufficiently explanatorily disciplined. Pinillos (2015, pp. 327–330) presses other worries. But I won't try to sort these worries out. Instead, I'll just declare:

- this ambitious appeal to merely possible discourses

to be another element of Fine's thinking that I'll mostly ignore. I will point out, though, that similar proposals are made by Taschek (see notes 48 and 49 above), and in different ways by Richard (the "third level" of his 1986–1987) and even Pinillos.[50]

My own reaction to the complaints about (1)/(22) and (23a/b) is to acknowledge that the novel semantic structure that Fine and/or Richard advocate may not be a cure-all for every challenge to Millianism. There may be sufficient reason to admit it nonetheless.

I won't attempt a systematic overview of other complaints that have been raised against Fine's account. But the papers by Soames cited in note 47 above deserve special mention, for two reasons. First, many of the complaints in these papers overlap ones in Soames (1987b), demonstrating shared challenges faced by Fine and by Richard. Second, some of Soames's complaints that we haven't yet discussed can be usefully framed against the way I've expositied Fine's account here.

One assumption behind some of Soames's arguments (such as 1987b, p. 113) is that attitude verbs will be truth-conditionally sensitive *at most* to coordination patterns local to their own complements. Bracketing the question of Soames's *entitlement* to that assumption (in 1987, before his opponents' views were elaborated), I observe that it is one that Richard and Fine resist: see move F4 and note 42 above. Admittedly, Fine does not give us any compositional semantics where this assumption is withheld. Such a semantics *can* be given (at least, if we suppress Fine's qualms about semantics being too "typographic," and also help ourselves to the machinery of "dynamic semantics"). But I won't have the space to give it here. Our discussion in §7 below will also fail to vindicate the ambitions behind F4.

Another assumption behind some of Soames's arguments (1987b, pp. 116–117; 1994, pp. 130ff.; 2010, pp. 473–474; 2012, pp. 243–244, 247–249, 250–251) is that attitude verbs on his opponents' accounts *will be* truth-conditionally sensitive to coordination patterns in their complements. That is, he assumes claim F3 above. But in fact, the full account offered by Fine permits attitude verbs to sometimes (on their "pure *de re*" reading) ignore those patterns. (We'll discuss this further in §6 below; see also the `value` operator I'll introduce in §7. As mentioned in note 43 above, Soames complains about this aspect of Fine's account. His discussion raises serious challenges that I won't attempt here to address.)

5

There are many differences in detail and in declared motivation between Richard and Fine.

One difference I'll point out is their approach to cyclic thoughts. Richard simply assumes we think with Mentalese sentences, and handles such puzzles accordingly. In Pryor (2016), I sort-of follow him, though at a greater level of abstraction: I appeal to the graph-theoretic structure of a subject's attitudes, rather than to any *concrete linguistic implementation* of that structure. Fine by contrast wants an explanation at the level of the *semantic content* of our thoughts, and so has to expand our conception of that content. Essentially, he can be understood as advocating that we count graph-theoretically distinctive relations among our attitudes as hitherto unacknowledged aspects of content. As I said before, Fine also wants his picture of the content of thought to be unified with his semantics for natural language.

Another difference is that Richard's account of the semantic differences between simple sentences like (1) and (5) – namely, they express different English-annotated RAMs – isn't "relational" in Fine's sense, but is instead of a kind that Fine will reject as too "typographic." Its advantage is that it enables Richard (unlike Fine, see point F1 above) to give a compositional (albeit context-sensitive) semantics for attitude reports. Given current technology, it's not clear how to offer such a semantics without provoking Fine's qualms about being too "typographic," or encumbered with conventional artifacts.

Despite these differences, there is also a good deal of shared ambition and predictions between Richard's and Fine's accounts. Fundamentally, both aim to explain how (1) and (5) can differ semantically, even when it's stipulated that `Cicero` and `Tully` don't differ in respects relevant to the content of isolated unembedded sentences like (23ab), where they

occur singly. When it comes to attitude reports, Richard can join Fine in denying that it's anything "intrinsic" to the content of the complement that constrains what thought is being attributed (Richard, 1990, p. 135) – though they develop this denial in different ways. Finally, the overlap in Soames's criticisms of them (mentioned at the end of the previous section) attests to the similar dialectical terrain they occupy.

In addition to the threads in Kaplan and Geach identified in note 14, other work that anticipates Fine's proposals – and does so especially closely – is Taschek (1992; 1995; 1998). I've already mentioned some points of contact in notes 44, 48, and 49; another is their shared rejection of a Compositionality Principle that tracks only the semantic contributions expressions make when they occur singly. An important respect in which Taschek diverges from Fine is that he takes the difference between (1) and (5) to merely be one of "logical potential," not to be part of their "information content."[51]

There is a range of other work since the late 1980s that has points of contact with Richard's and Fine's accounts. The other authors cannot be understood to defend a common position or thesis, but in each case there is some clear affinity with the Putnamian idea that guided Richard and Fine.

Some of this other work focuses on cyclic thoughts. For example, Campbell (1987–1988) talks of an inference "presuming" or "trading on" identity, without the subject needing any *identity judgment* or premise. Lawlor (2002) talks of "anaphoric thinking." See also Perry (1980)'s discussion of "internal identity," and work in the "mental files" tradition like Recanati (1993; 2012, esp. chs 8–9; 2013, esp. §4 which replies to Goodsell, 2013; and Recanati, 2015, §I).[52]

Other work does focus on *de jure* codesignation in language; however, some of the authors deny that the phenomenon makes a *semantic* difference. See for example the work by Schroeter cited in note 41. Taylor (2003b; 2015) contrasts "explicit" versus "coincidental" codesignation, and says that the former is "syntactically or structurally" marked. These differ in what they "manifest" or "display," and in their "dialectical significance," but not in their contributions to propositional content (see esp. Taylor 2003b, §§2 and 8; 2015, pp. 257–259). As explained above, Taschek (1992) and Richard (1986–1987) also hesitate to make such differences be semantically encoded.

Unlike them, Pinillos (2011; 2015) *does* claim that the phenomenon he's exploring makes a semantic difference. He uses the phrase "*de jure* coreference," meaning by it something narrower than I express with "*de jure* codesignation," since he says bound variables don't refer and so never corefer (2011, pp. 318–320, but see also his n. 10). But the underlying semantic mechanism Pinillos is studying is one he thinks *does* apply to bound variables, and to empty names (2011, pp. 317–318; 2015, pp. 324–325), as well as to ordinary names and to anaphoric but unbound pronouns, like `his` in the "strict" reading of (4a). Pinillos calls this mechanism "p-linking" (2011, pp. 317ff.). Though his terminology may suggest otherwise, Fine also wants his notion of "semantically required" or "strict" coreference to apply to bound variables and to empty names (see notes 39 and 35 above). One fundamental difference between Fine and Pinillos is that the former takes his phenomenon to be an equivalence relation, at least when we confine ourselves to intrapersonal cases (Fine, 2007, pp. 55–56, but see note 44 above). It is central to Pinillos's account, on the other hand, that the underlying semantic mechanism can be intransitive and non-Euclidean even in that domain. He gives examples like the following:

(24) `As a matter of fact, the orator Cicero is my brother Marcus; you will get to meet the great Marcus Tullius Cicero tonight.`

Pinillos would claim that the first and third name occurrences, and also the second and third, are p-linked, but the first and second are not.[53]

Pinillos states several "axioms" governing the behavior of p-linking. Some of these govern the relation between it and coreference, or between it and variable binding (Pinillos endorses a claim like F2 from §4 above). Most interesting is his Axiom 2 (2011, p. 318), which Pinillos understands to have something like F3 as a consequence. This is closely related to Pinillos's Principle of Attitude Closure, which he takes to be partly criterial for the phenomena he's exploring. This Principle says that when the complement of a belief report exemplifies *de jure* coreference, that report will entail a corresponding report using the complement's existential generalization. For example:

(5*) Anita believes that Cicero admired Cicero.

will entail:

(10*) Anita believes that someone admired himself.

Pinillos says this Principle may hold for more attitudes than just belief, though not for every attitude.

We'll discuss F3 and Pinillos's Attitude Closure in more detail in the next section.

Pinillos's other criteria for the phenomena he's exploring are principles that are reminiscent of Fine's claims about the epistemology of language use. Pinillos says that anyone who fully understands a sentence exemplifying *de jure* coreference will be in a position to infer its existential generalization, for example, from (5) to:

(10) Someone admired himself.

or from (3d) to:

(25) Terentia betrayed someone who was unhappy.

Pinillos says also that anyone who fully understands a sentence exemplifying *de jure* coreference will know of the relevant expressions that if they both manage to designate, they designate the same thing.[54]

My own stance towards these "criteria" is very guarded. Because I have different background views about the epistemology of language use than Fine and Pinillos – which dispute seems like it should be orthogonal to the issues we're exploring – I'm inclined to doubt that any linguistic properties meet the epistemic tests they propose. Perhaps refined and more qualified versions of these tests will prove more acceptable. But even in advance of knowing what those will look like, I do still find myself sympathetic to the possibility of a genuine semantic phenomenon in the neighborhood they're exploring. So my preference is to detach the epistemic criteria from the semantic proposals.

Goodsell (2014) shares some of my qualms about the epistemic commitments of Pinillos's criteria. She also complains that these criteria wouldn't be able to characterize *de jure* codesignation across multiple speakers or in non-declarative constructions (p. 298); also that, as Pinillos states them, the criteria don't sufficiently respect the fact that one can understand a belief report without oneself, or the report's subject, knowing that all its constituent expressions refer (pp. 304–305).

Finally, I'll note that Fiengo and May (2006) has similar motivation and points of contact with Richard's and Fine's views. But it also differs in complex ways that I can't explore here. (Additionally, Fiengo and May embrace some of the controversial views about the epistemology of language that I prefer to separate from the semantic proposals we're considering.)

6

Claim F3 and Pinillos's Attitude Closure have come up several times in our discussion. The status of these claims is contested.

Suppose Diego has a long dream in which, contrary to fact, he's married. Inside the dream, he often finds himself waking to the sight of scattered socks, and supposes this may have been done by a mouse. Being terrified of mice, he begins a nightly campaign to wake his more fearless wife so she can catch the sock-mover. But they never encounter a mouse. Finally, it emerges that Diego and his wife have all along been moving the socks themselves, while sleepwalking. Thus ends Diego's dream. Diego, now awake for real, reports his dreamt campaign with bemusement:

(26) `All along, I was hoping my wife and I would catch ourselves!`

The pronoun `ourselves` gets its reference from `my wife and I`, so these expressions should be *de jure* codesignative. But then the acceptability of the report would conflict with any claim like F3. In the scenario described, it was *not* the case that Diego all along bore a hopeful attitude that represented the chasers and the agents chased as one, and from whose aim he could infer `Someone catches themself`.

This example also puts pressure against Salmon's and Soames's views, summarized in §1 above. The fact that Diego has no real wife suggests that `my wife and I` occurs inside the report's complement, as it appears to, rather than taking wide scope. (The reflexive morphology of `ourselves` also suggests this.) So on a bound-variable analysis of referential dependency, the report has the form:

(26′) `All along I was hoping: (λx. x would catch x) (my wife and I)`

And on Salmon's and Soames's accounts, any such report would *also* attribute a hope that represents the chasers and the agents chased as one.

Soames several times appeals to similar examples, with the form:

(27a) `Juan told Maria that he wasn't Juan/that man/him.`[55]

In all its variations, this is an acceptable report of a scenario where Juan attempted to mislead Maria about his identity: that is, he said to her something like `I am not Juan`, or `Juan is shorter than me`. Another of Soames's examples has the form:

(27b) `Each man told Maria that he wasn't that man.`[56]

Soames assumes that the singular terms in the complements of (27ab) should each be *de jure* codesignative with the subject of the report. Presumably, they will then be *de jure*

codesignative with each other; and if F3 is in place, they'll report Maria to have been told something from which she could infer `Someone is not himself` – contrary to Soames's description of the scenarios.[57]

Spelling this out more carefully, the assumptions that raise trouble are that:

(i) if two expressions are *de jure* codesignative with a third, they are also *de jure* codesignative with each other

(ii) if the expressions in the complement of (27ab) are codesignative with each other, those reports characterize Maria as having been told contents from which she could infer `Someone is not himself`. This is the claim I earlier labeled F3.

Some advocates of *de jure* codesignation deny that (ii/F3) always holds: see for example Recanati (2012, §9.2) and Goodsell (2014, pp. 302–304). Fine himself only endorses F3 for some readings of attitude reports.

Pinillos acknowledges examples like (27ab) (2011, pp. 315–316 and n. 41). Instead of rejecting (ii/F3), he prefers to reject assumption (i). He denies that *de jure* coreference (or p-linking) is an equivalence relation; so the singular terms in these reports' complements needn't be p-linked with each other, after all.

Like Pinillos, Soames also sees these examples as demonstrating a failure of Euclideaness and/or transitivity. He thinks we should see the referential dependencies in (27b) and some versions of (27a) as having the form:

(27c) `_____ told Maria that he wasn't that man.`

This contrasts with other examples where referential dependencies have the form:

(28a) `Cicero expected that he would eventually admire himself.`

In (28a), `himself` can't be directly referentially dependent on `Cicero`, since if it were, it could not have reflexive morphology:

(28b) `Cicero expected that Terentia would eventually admire *himself/him.`

The main difference between Soames and Pinillos is that only the latter says these referential dependencies will always be semantically encoded. As I explained in §1 above, Soames thinks the dependencies show up in the semantics iff they're generated by variable binding. That will only partially overlap with Pinillos's ambitions for the notion of *de jure* codesignation.

Higginbotham and Heim also argued for a distinction in dependency patterns like (27c) versus (28a). Their examples have an advantage over Soames's, in that his case for denying `that man` is dependent on `he` in (27ab) turns on assumptions like (ii/F3) above (or, on his own view, on (Z2) and (Z3) from §1). Since the status of these assumptions is unclear, demonstrations that don't need them are more forceful.

Heim (1998, examples 26a–g and 27a–d) gives contrasts like these:

(29a) `Every student claimed that he called his mother before the teacher did.`

(29b) `Every student claimed that he called his mother before the teacher did.`

On Heim's view, the referential dependencies exhibited in (29a) invite the "strict" reading, where it's the teacher's calling *the students*' mothers that is being discussed. The referential dependencies in (29b) invite the "sloppy" reading, where it's the teacher's calling *the teacher's own* mother that is being discussed.

Higginbotham (1985, examples 68–74) gives contrasts like these:

(30a) `They told each other they would succeed.`

(30b) `They told each other they would succeed.`

On his view,[58] the referential dependencies exhibited in (30a) describe scenarios where the subjects said to each other `I will succeed` or `We will succeed.` The referential dependencies in (30b) describe a scenario where they said `You will succeed`, as is also unambiguously reported by:

(30c) `They told each other to succeed.`

However, though Heim and Higginbotham argue that these dependency patterns should be distinguished, they do not argue for their being semantically encoded. So far as I can see, their discussions just cited could be developed along the lines of either the Salmon/Soames stance from §1 above, or the Richard/Fine/Pinillos stance. (The former will have been more familiar to Heim and Higginbotham.)

7

We can get further insight into the Richard/Fine/Pinillos proposals by considering some ideas from computer programming. One obstacle to this is that many readers will have little fluency with the kind of programming we need. But we can work around that. Those readers who are acquainted with programming may be thinking of it on the model of a sequence of instructions. That is the "imperatival" model of computation. But there are other models, too, such as the "declarative" or "functional" model. The existence of this model of computation may be less familiar, but *what it understands computation as* is closer to ideas that philosophers and linguists are well-acquainted with, like formulas of predicate logic. On the declarative model, arithmetic expressions like $2 + 7 \leq 9$, $2 + 7$, and 2 all count as programs, which evaluate to true, 9, and 2, respectively. The set of inequalities $\{2 + x \leq 9, x > 5\}$ might be a program that evaluates to a set of assignments binding x to 6 or 7. The regular expression `/ima[a-z]*ng/` might be a program that evaluates to the set of English words {`imaging`, `imagining`}. And so on. There's nothing in either of these models of computation that prevents it from expressing or achieving all that the other can; their styles of doing so will just tend to be different. We'll work with the declarative model.

If readers have ever encountered a formal semantics for a programming language (from either of the models of computing just described, or another), it's likely to be of the form computer scientists call "operational semantics." This will have surface resemblance to a

proof theory, or a system of rules for rewriting the program text. (Those analogies have defects, but that's not important for our purposes here.) But there are other forms of semantics for programming languages. The one most similar to what philosophers and linguists think of as semantics is called "denotational semantics." If you've ever seen a denotational semantics for even a simple (yet Turing complete) programming language, you will have been struck by how much more mathematically sophisticated it is than anything appealed to in natural language semantics.

All of this – associating "programming" with the imperatival model of computation, thinking of formal semantics for programming languages in operational rather than denotational terms, and the sophistication of actual denotational semantics – can encourage the impression that the semantics of programming languages is profoundly different from, and will teach us little about, what we understand as semantics for languages like English (or Esperanto, or arithmetic).

However, I invite you to consider some notions from declarative programming that I will structure as a series of extensions to (a decidable fragment of) arithmetic. I will present these in a way that aims to maximize their familiarity. I won't present a formal semantics for these notions here. The extensions to arithmetic that one sees in a full-blown declarative programming language fall into "lightweight," "middleweight," and "heavyweight" categories. The "lightweight" additions don't require anything fundamentally new for their semantics. The "middleweight" additions do; but no more radically new than philosophers and linguists are already familiar with from Lewis (1979) and discussions of "dynamic semantics." The "heavyweight" additions include things like unlimited recursion, and giving a denotational semantics for these does require sophisticated techniques. But none of what we'll be doing here needs to make use of those notions. We can limit ourselves only to "lightweight" and a few "middleweight" additions.

Let's begin with restricted quantification, with which I assume familiarity. Throughout I'll use *x*, *y*, … as variables ranging over the domain ℕ. Later I'll use *xs*, *ys*, … as variables ranging over lists of ℕ, and *f*, *g*, … as variables ranging over functions of one or two arguments (usually mapping ℕs to ℕs, but sometimes with other types).

The sentences:

(31a) $\forall x < 7 \left(2 + x \leq 9\right)$

(31b) $\forall x = 7 \left(2 + x \leq 9\right)$

express truths, given the standard interpretations of <, ≤, =, +, and the numerical constants. We'll represent that by writing:

(31b) $\forall x = 7 \left(2 + x \leq 9\right) \quad \Rightarrow \quad$ true

Our ⇒ notation also permits us to state what values sub-sentential expressions have, on the assumed interpretation. For example:

(32a) $2 + 7 \quad \Rightarrow \quad 9$

and:

(32b) $2 + x \quad \Rightarrow \quad 9$

on an assignment that binds variable *x* to the value 7.

An interesting difference between (31a) and (31b) is that the latter uses a restrictor that we know in advance is satisfied by exactly one value in the relevant domain. It will be useful to introduce a special syntax for this case:

(31c) `let` $x = 7$ $(2 + x \leq 9)$ ⇒ true

A natural extension of this syntax permits us to partially specify our assignments in the case of sub-sentential expressions, too, like (32b):

(32c) `let` $x = 7$ $(2 + x)$ ⇒ 9

We can specify assignments more fully by embedding uses of `let`:

(33a) `let` $x = 3$ (`let` $y = 4$ $(2 + x + y)$) ⇒ 9

which we might abbreviate as:

(33b) `let` $x = 3, y = 4$ $(2 + x + y)$ ⇒ 9

Now let's see how to express the same things using the syntax of the programming language Scheme.[59] This is how you write $2 + x + y$ in Scheme:

(34)
```
(+ 2 x y)
```

and this is how you write the equivalent of (33ab):

(33c)
```
(let ([x 3][y 4])
     (+ 2 x y))
```
⇒ 9

For later comparison, it will be helpful to note that just as you can say $\forall x\, \exists x\, (x = 9)$ in predicate logic, and inside the parentheses only the innermost effect on the binding of x will be operative, so too can you say:

(35a)
```
(let ([x 0][y 0])
     (let ([x 3][y 4])
          (+ 2 x y)))
```
⇒ 9

How this is described in programming circles is that the innermost `let` introduces new local bindings for x and y that "shadow" their more global bindings. These local bindings are in place only until the parenthesis that closes the innermost `let` expression.

(35b)
```
(let ([x 0][y 0])
     ((let ([x 3][y 4])+) 2 x y))
```
⇒ 2

What happened here is that the local bindings to x and y were in place only for the interpretation of the symbol + (where they made no difference). When we evaluate the operands of +, we use the more global bindings for x and y.

The Scheme programming model countenances several types of simple and compound values. Simple values include things like numbers, truth-values, and various functions; we won't concern ourselves with other simple values. Compound values include things like lists of other values; we won't concern ourselves with other compound values. In Scheme, a list can contain a variety of values, such as numbers, truth-values, and yet other (embedded) lists. But we will mostly work with homogeneous lists of numbers. Here is one way to build a list in Scheme:

```
(36a)  (list 2 3 4)                          ⇒  (2 3 4)
(36b)  (let ([x 3] [y 4])
           (list 2 x y))                     ⇒  (2 3 4)
```

The syntax of Scheme includes *keywords* like `let`; simple expressions including *symbols* like `+`, *x*, and *y*, and *literals* like the numerical constants and `true` and `false`; and compound expressions which are lists of other expressions. Scheme *evaluates* symbols to the values those symbols are bound to, at that position and stage in the program. It evaluates literals to themselves. It evaluates lists of expressions by evaluating the head of the list, and if the result is a function, applying it to the argument values that the rest of the list evaluates to. (If the head of the list doesn't evaluate to a function, the list is not evaluable.) Lists beginning with keywords are evaluated in special ways.

Here are some more operations on lists. Note that here we bind a variable to a list rather than to a simple value. I'll follow the convention of using plural variable names like *xs* for this.

```
(37a)  (let ([xs (list 2 3 4)])
           (head xs))                        ⇒      2
(37b)  (let ([xs (list 2 3 4)])
           (tail xs))                        ⇒    (3 4)
(37c)  (cons 2 (list 3 4))                   ⇒  (2 3 4)
```

The `cons` function is so-named because it *cons*tructs a list, from a specified head element and rest of the list. This should now make sense:

```
(38)   (let ([xs (list 2 3 4)])
            [ys (cons 1 (tail xs))])
           ys)                               ⇒  (1 3 4)
```

The symbol `+` comes pre-bound to the standard addition function. Scheme also lets you build custom functions, like so:

```
(39)   (λ (x y) (+ 2 x y))
```

This evaluates to a function that accepts two arguments which it then adds to 2, and returns the result. (Scheme doesn't have any canonical way to display this function, so I have omitted the ⇒.) We can apply that function to values like so:

```
(40a)  ((λ (x y) (+ 2 x y)) 3 4)             ⇒  9
```

This behaves the same as (33c). We can also bind variables to functions that we build in this way:

```
(40b)  (let ([g (λ (x y) (+ 2 x y))])
            (g 3 4))                                  ⇒  9
(40c)  (let ([g (λ (x y) (+ 2 x y))]
             [x 3])
            (g x 4))                                  ⇒  9
```

λ-expressions can contain variables that are bound outside of them (variables that are "locally free"):

```
(41)   (let ([y 4]
             [f (λ (x) (+ 2 x y))])
            (list (f 3) (f 13)))                      ⇒  (9 19)
```

These free variables are understood in what programmers call the "lexical" rather than the "dynamic" style. That is:

```
(42)   (let ([y 4]
             [f (λ (x) (+ 2 x y))]
             [y 0])
            (list y (f 3) (f 13) y))                  ⇒  (0 9 19 0)
```

The innermost binding of *y* to 0 affects the interpretation of that variable in the final list expression; however, it doesn't affect the interpretation of *y* inside the definition of the applied function *f*. The other understanding, where the result of the above would be (0 5 15 0), is also coherent; but it's not what Scheme uses here.

Contrast (42) to:

```
(43)   (let ([y 4]
             [f (λ (x) (let ([y 0]) (+ 2 x y)))])
            (list y (f 3) (f 13) y))                  ⇒  (4 5 15 4)
```

Here the more local binding for *y* is inside the definition of *f*, so it *does* affect the results we get when applying *f*; also, that binding is in effect only *inside* the definition of *f*, so it *doesn't* affect the interpretation of *y* in the final list expression.

This finishes our introduction to (the core of) the "purely declarative" part of Scheme. The ideas exhibited here all fall under the heading of what I called "lightweight" additions to arithmetic. Next we move on to some "middleweight" additions.

Lists of the sort we've worked with so far are *static* values. Once built, such a value never changes. In this respect they are just like numbers and truth-values. However, Scheme also has a notion of a *mutable* list, which can at different stages of its existence contain different elements. We can build and mutate such a list like this:

```
(44)   (let ([xs (mlist 2 3 4)]
             [_ (set-head! xs 1)])
            xs)                                       ⇒  (1 3 4)
```

A few comments about this. First, note the different syntax: `mlist` rather than `list`. Second, this example contrasts with (38) where we bound *ys* to a version of *xs* with a new head, but *xs* itself would still have evaluated to the list (2 3 4). Third, the usefulness of `set-head!` is in the *side-effect* you get from applying it, not from its return value. In most Schemes, `set-head!` doesn't return anything useful, so we assign its return value to the "throwaway" variable _. Fourth, in introducing such values and operations into the language, we've abandoned what programmers call "referential transparency." This loosely coincides with what philosophers mean by that phrase. Consider:

```
(45)  (let ([f (let ([ys (mlist 0)])
                 (λ (x) (let ([y  (head ys)]
                              [_ (set-head! ys (+ 1 y))])
                          (+ 2 x y))))])
        (list (f 3) (f 3) (f 3)))                          ⇒ (5 6 7)
```

If you reason through the evaluation of this expression, you'll see that at each application of *f* to the constant argument 3, we end up adding a different head-value from the changing list *ys*. So the result of applying *f* is not determined by the values supplied to it as arguments.[60]

Scheme also has some built-in functions that return truth-values as their result. For example:

```
(46a) (let ([y 4])
        (equal? y 3))                                      ⇒ false
(46b) (let ([ys (list 2 3 4)])
        (equal? ys (list 1 3 4)))                          ⇒ false
```

Instead of `equal?` we could also have written `egal?`. With non-mutable values, there is no difference between these. But with mutable lists, we face a choice. Two variables might be bound to distinct mutable lists, which happen currently to contain the same elements. A mutation to one of the lists would not in itself have any effect on the other list. Should we count such lists as equal, since they now contain the same elements? Or should we say that they're unequal, since they are distinct, independently mutable containers? Scheme handles this by saying such lists are `equal?` but not `egal?`. Thus:

```
(47)  (let ([xs (mlist 2 3 4)]
            [ys (cons 1 (tail xs))]
            [zs xs]
            [_ (set-head! xs 1)])
        (list zs (egal? xs zs) (egal? xs ys) (equal? xs ys)))
                                                   ⇒ ((1 3 4) true false true)
```

We've been working with the notion of a mutable list *value*. But so far, our understanding of how *variables* work has stayed the same. Once a variable is bound to a value, it stays so bound. That binding may perhaps be "shadowed" by a more local use of the same variable symbol, but it continues to exist and will again be visible after the syntactic scope of the more local expression expires. But in fact, Scheme also has a notion of *mutable variable bindings*. It allows us to say:

```
(48)  (let ([y 4]
            [f (λ (x) (let ([_ (set! y 0)]) (+ 2 x y)))])
           (list y (f 3) (f 13) y))                    ⇒  (4 5 15 0)
```

Contrast (43), where the final element in the result was 4. Here we don't "shadow" the outermost binding of *y* with a new binding. Instead we *mutate* that very outermost binding. The result is that even *outside* the definition of *f*, after *f* has been applied, that new mutated binding will still be visible.

One of the reasons I chose Scheme for these examples is that it has *both* mutable values *and* mutable variable bindings. It's most common for programming languages to have only one of these. Here is an example of using both notions in Scheme:

```
(49)  (let ([xs (mlist 2 3 4)]
            [zs xs)
            [_(set-head! xs 1)]
            [_(set! xs (list 0 0 0))])
           (list xs zs))                               ⇒  ((0 0 0) (1 3 4))
```

In (47), we used the variable *xs* to mutate the *single list value* that both *xs* and *zs* were bound to; then when we evaluated *zs,* this change was visible. The same is true in (49); but after mutating that list value, we next mutate the *variable xs* to become bound to a new list. *zs* still stays bound to the old list.

At this point, the imaginative reader who's read the rest of this chapter will wonder whether there might be a way to make variables be even more intimately related than *xs* and *zs* were in the above examples: to not merely happen at one stage in the program to designate or be bound to the same value, but to be made *de jure* codesignative, even through mutations of either variable's binding.

The answer is that yes, this is possible, but it requires going beyond the most familiar features of Scheme.[61]

Consider:

```
(50)  (let ([xs (list 2 3 4)]
            [zs xs])
           (alias ([ws xs])
                  (let ([_ (set! xs (list 0 0 0))])
                       (list xs ws zs))))              ⇒  ((0 0 0) (0 0 0) (2 3 4))
```

Here, as before *zs* is merely `egal?` to *xs*, so when we mutate the binding of the latter, the former still says bound to the original list. However, we have used the new keyword `alias` to make *ws* be so intimately related to *xs* that changes to the binding of either of them also affect the other. Aside from the `set!` operation, none of the other language features we've seen (so far) differentiate between `alias` and `let`.

We can also introduce an operation that works like λ, except that it accepts only variables as operands, and it associates the parameters in its definition with those operands using `alias` rather than `let`:

```
(51)  (let ([f (aliasλ (ws) (let ([_ (set! ws (list 0 0 0))])
                              ws))]
            [xs (list 2 3 4)])
           (list xs (f xs) xs))                    ⇒ ((2 3 4) (0 0 0) (0 0 0))
```

Because of how the parameter *ws* internal to the definition of *f* gets associated with the variable *xs* when we evaluate (`f xs`), mutating *ws*'s binding also ends up mutating the variable *xs*.

With these resources in hand, we can define a third kind of equality predicate:

```
(52)  (let ([aliased? (aliasλ (u v)
                              (let ([u0 u]
                                    [_ (set! u (mlist 0))]
                                    [result (egal? u v)]
                                    [_ (set! u u0)])
                                result))]
            [x 3]
            [z x])
           (alias ([w x])
                  (list (egal? z x) (aliased? z x)
                        (aliased? x x) (aliased? w x))))
                                                   ⇒  (true false true true)
```

This makes use of the expression (`mlist 0`), which creates a new mutable list. Such a value is guaranteed to not be `egal?` to any other value that variables are already bound to. What this definition of `aliased?` does is save the original value of its parameter *u* as u_0, then mutate the binding of parameter *u* and see whether the other parameter *v* ends up being `egal?` to the new value. It saves the result of that test as *result*, then restores *u* to its original value (in case the invoking context needed it), and returns *result*.

This `aliased?` predicate is like *f* in example (45), in that its result is not determined merely by the argument values passed to it (*z*, *x*, and *w* all have the same value, 3). But `aliased?` is more interesting, in that it *does* track whether its operands stand in the *de jure* codesignating relation that different occurrences of *x* (in the same binding context) stand in to each other, and that *w* and *x* also stand in. Thus the last two tests in the final list expression evaluate to true; whereas *z* on the other hand is merely `egal?` to *x*, not also `aliased?`

Defining `aliased?` in the way just demonstrated makes essential use of `(mlist 0)` and `set!`, which involve the two kinds of mutation we explained above. This is a good way to get an intuitive handle on how such a predicate could even be possible. But it may give the impression that in less "dynamic" languages, that don't have any of these mutation novelties, a notion like `aliased?` would be inexpressible. To counter that impression, I'll report that it's possible to define `aliased?` in many implementations of Scheme in a way that doesn't make use of mutable values or bindings. The mechanisms required to do this are complex; see the URL in note 59 for details. Regardless of how feasible it is to implement `aliased?` in these specific programming languages, it could always be given a semantics as a native programming idiom even without the language also having mutable values and bindings. (But the strategies for giving a formal semantics for these several notions will be similar.)

It should be possible to use `aliased?`, `alias`, and `aliasλ` all together. We demonstrated how the first and second should work together in (52). Here is how they should interact with the third:

```
(53)  (let ([x 3]
            [g (aliasλ (u v)
                       (list (aliased? u u) (aliased? u v)
                             (aliased? u x)))]
            [z x])
           (alias ([w x])
                  (list (g x x) (g w x) (g z x))))
                       ⇒    ((true true true) (true true true) (true false false))
```

Because *w* and *x* are aliased, and passing them to *g* aliases them to *g*'s internal parameters, all of the tests in *g* come out true when any combination of *w* and *x* are supplied as operands. But *z* merely happens to currently be bound to the same value as the other variables. It isn't `aliased?` to them. (Of course, *z* is `aliased?` to itself.)

The predicate `aliased?` is not (or not merely) a quotative operator, as it doesn't just test for whether its operands are syntactically matching. Rather, it's sensitive to the patterns of *de jure* codesignation in those operands. It provides an example of the kind of predicate that we observed Richard offering (and Fine at least gesturing at). (But it doesn't have the interesting behavior I labeled F4 in §4 above.) These authors think a linguistic context like `Anita believes that _____ admires _____` may evaluate to true when `Cicero` and `Cicero` are supplied as operands, but false when `Cicero` and `Tully` are supplied – even if those names themselves have the same semantic value. In the same way, (`aliased? x x`) may be true but (`aliased? x z`) false, despite *x* and *z* being bound to the same value.

Here is another connection to our earlier discussion. In the following:

```
(54)  (let ([value (λ (y) y)]
            [x 0])
           (list (aliased? x x) (aliased? x (value x))))
                                                    ⇒    (true false)
```

we define an operation `value` that extracts the value that *x* is currently bound to, returning it in a form that doesn't count as `aliased?` to *x* (or any other variable). We might then think of contexts like `Juan told Maria that _____ wasn't _____` as being false when the *de jure* codesignative expression-occurrences `he` and `him` are supplied as operands, but as being true when `he` and (`value him`) are supplied (see note 57). Perhaps the key to understanding Soames's (27a) is to interpolate such an (unpronounced) operation. (This is similar to Fine's strategy of positing a special *de re* reading of the report, and vulnerable to some of the same objections.)

If one is new to the programming idioms explained here, what we've walked through may seem exotic and to shed little light on the intelligibility or prospects of predicates with `aliased?`-like behavior in natural language. However, as I said at the start of this section, there is little in the "lightweight" additions that one should find semantically novel. And of the "middleweight" additions, we made use of mutable values and variables only as a stepping-stone to get to an understanding of how `aliased?` and its partners work. In principle, as I said, we don't need to define them with mutation. They can exist in languages that are mutation-free.

As I acknowledged earlier, programming languages can and often do make use of genuinely exotic ideas that require more sophisticated semantic techniques than readers of this chapter may be familiar with. But none of those ideas – unlimited recursion, continuations, fancy types – are needed to make sense of `aliased?` and its friends.

I hope this section contributes towards "domesticating" the kind of predicates that fans of *de jure* codesignation are friendly to, that is, making them seem less alien and somewhat less "magical." As we've seen, such predicates are needed anyway for some formal languages, ones that ought not seem miles away from natural language. In my view, the work summarized in this chapter's earlier sections creates a case for exploring whether this kind of semantic structure is present in natural language too: perhaps in ways that encode a semantic difference between demonstrative and the "strict" anaphoric uses of pronouns from §1, or perhaps in ways that affect the truth-conditions of attitude reports, or perhaps in ways that encode the different patterns of anaphora posited by Heim and Higginbotham in §6 for unembedded sentences.[62]

Notes

1 Compare Higginbotham (1985, examples 62–63); Heim (1998, pp. 213–216, 241); and McKay (1991, p. 718). On the contrast between merely intended coreference and the referential dependence we see in examples that follow, see Evans (1980, §5), Recanati (2012, §8.1), and Fiengo and May (1994, esp. pp. 3–13, 49–51). Goodsell (2014) articulates a notion of "assumed coreference," also distinguished from the relations exemplified below.

2 When talking about this example, I'll rely on the reader to understand that the sentence is being used and understood in the way described; and similarly for later examples. Of course, the *displayed sentence* might also be used in the earlier way. It will simplify exposition for me to presume (*contra* Lasnik, Bach, and some other authors discussed below) that the demonstrative and anaphoric uses of pronouns are syntactically different.

3 For a survey of some accounts, see Dahl (1973). The "bound-variable" analysis described below is proposed in Keenan (1971), Williams (1977), Reinhart (1983, ch. 7), and elsewhere. For convenience, I will apply the labels "sloppy" and "strict" also to the hypothesized disambiguations of sentences like (3e).

4 This is sometimes denied, for instance by Recanati (2005, p. 308) and Salmon (1986b, pp. 50–52), though Salmon (1992, pp. 59, 65) allows that (3b) may be ambiguous. The "strict" reading is attested by Dahl (1973, p. 84) and Evans (1977, p. 95), who also cites Dummett and Partee. See Partee (1989, n. 4). The "strict" reading will dominate if we replace `Atticus` with a female name like `Terentia`. In that case, it's controversial whether the "sloppy" reflexive reading would be available at all.

   Sometimes it's suggested that `Cicero mourned his own daughter` gives us a way to force the "sloppy" reading. But McKay (1991, n. 21) convincingly shows that "strict" readings are still possible with such sentences. See also Reinhart (1983, ch. 7, n. 13).

5 Bach is in one way less radical than Lasnik, as he allows that expressions like `himself` and `his own` may have the distinctive syntax and/or semantics that other theorists attribute to our (3) examples more broadly. In another way, Bach is more radical, as he argues that pronouns apparently bound by quantifier phrases should be understood on the model of (2ab), too. See also Recanati (2005).

6 Compare Tomioka (1999, examples 7 and 8, and n. 4; also examples 23, 24, 27). For further discussion, see Soames (1989–1990, pp. 211–212), McKay (1991, pp. 721–723 and n. 30), Salmon (1992, pp. 65–66; 2006a), and Soames (1994, pp. 120ff.).

7 See Salmon (1986b; 1992; 1989, pp. 216ff.; 2006b; 2010, pp. 447–449); Soames (1985, n. 12; 1987a, n. 24 and pp. 58–61; 1987b, esp. §8; 1989–1990, pp. 204ff.). Salmon counts the propositional

forms in (Z2) as "logically equivalent" (2010, p. 451; also 1986b, pp. 51–55 and n. 23), but not synonymous, and he denies that ($\lambda x.$ *x* admired *x*) *a* will always be *inferable* in the sense we're considering from the others.

At (2010, pp. 453–454, 456–457), Salmon discusses whether we should count λ-conversions as synonymy-preserving when the bound formula is monadic.

Other endorsements of parts of the stance I attribute to Salmon and Soames include Kaplan (1986, pp. 269–271), Higginbotham (1991, pp. 360–361), and McKay (1991). See also note 3 above; and Bach (1987, pp. 256–257), though the rest of Bach's chapters 11–12 argues against extending the view beyond pronouns like `himself`. Salmon also cites Wiggins (1976b, pp. 164–166; 1976a, pp. 230–231).

8 This is meant to suggest the image of their structure somehow looping back on itself. Compare the idiom `(cons #1=expr #1#)` in some implementations of the programming language Scheme, which can be used to specify cyclic data structures.

Some authors already call these "reflexive" contents – but we've used that term above in other senses: for the reflexive morphology of `himself`, and for the "sloppy" reflexive readings of some of our examples. Salmon and Soames would also use the term for bound predicates with the structure of (3b′), and for our present notion. But since it's controversial how much these different phenomena coincide, it's conceptually more hygienic to introduce a new label here. (There is also the mathematician's notion of a "reflexive relation," which is not directly relevant to what we're discussing.)

Fine calls cyclic contents "coordinated," but I suspect this term is too closely associated with his own views. (And Fine also uses the term to refer to a kind of linguistic structure, as well as a kind of content.)

9 Fine says only a few words about anaphoric pronouns (see note 33 below), but the bound-variable analysis does not seem to constitute his understanding of them. It's clear that it doesn't constitute his understanding of (5) (see Fine, 2007, pp. 69–70). It's not clear whether Fine regards (3b) and (5) as semantically equivalent. Soames (2010; 2012) interprets him as giving them different contents, albeit contents the understanding of which requires recognizing that it's true iff the other is. As explained above, Soames's own view is that (3b) does have the form (3b′), and (5) does not, and neither is (3b′) inferable from the content of (5). But on his view, standardly someone who utters (5) will *also thereby assert* the content (3b′) (see Soames, 1987b, §8; 2002, esp. chs 3 and 8; 2005). See also Richard (1990, pp. 216–217).

Soames proposes treating the "strict" readings of anaphoric (3) sentences semantically on a par with the (2) examples (1989–1990, pp. 210–211). See also Heim and Kratzer (1998, pp. 240–241).

10 See Pinillos (2011, example 18). I'm ignoring accounts of (11a) in terms of focal alternatives. On some variations of the account sketched here, the first conjunct of (11bc) would be presupposed rather than asserted. Similar examples could be given using `Even Cicero ...` or `It was Cicero who ...`

11 See Heim and Kratzer (1998, §9.3) for an introductory overview.

12 Compare Evans (1980, examples 49, 53, 55) and Heim (1998, examples 13, 20).

13 Church (1954) accepts this consequence, and argues that the doubt really available to these subjects is not about lawyers but rather about the *predicates* `lawyer` and `attorney`. But Church did not think attitudes were *in general* about linguistic expressions.

14 Kaplan's views were presented in unpublished lectures delivered in 1985 (reported in Soames, 1987b, p. 112 and n. 19; and Salmon, 1986a, p. 164), and also in a few lines in Kaplan (1990, n. 6). There are also passages where Kaplan embraces the competing Salmon and Soames view: see note 7 above.

Geach's views are complex. One challenge is that he denies that sentences always have a single "subject-predicate analysis" (even in the absence of syntactic ambiguity; see Geach, 1962/1980, §§24, 27; 1975b, pp. 144–146). At the same time, he claims that some pairs of sentences are "logically equivalent," "say the same thing," or "have the same import," and yet "contain different predicables" or "predicate different things" of their respective subjects (1962/1980, §§78, 80). It's

hard to know how we are to identify *the* predicate of a sentence in some of these cases. In some places, though, Geach explicitly affirms that sentences like (5) predicate the same thing of Cicero that `Atticus admired Atticus` (and `Atticus admired himself`) predicate of Atticus (1962/1980, §26; 1965, p. 112; 1975b, pp. 139–140, 141, 147). Given Geach's analysis of reflexive pronouns (1962/1980, §§80–84), this means he'd see them all as "containing" the predicate $\lambda x.\ x$ admired x. In other cases, though, he denies that recurring names induce this predicate structure. For example, `Only Satan pities Satan` differs in meaning from (the "sloppy" reading of) `Only Satan pities himself` (1962/1980, §80). Geach does acknowledge a "strict" reading of the similar sentence `Not only Socrates loves his wife`, but although he allows that it's "logically equivalent" to `Not only Socrates loves Socrates's wife`, this is one of the cases where he says the paired sentences have different analyses (1975a, pp. 197–198). He interprets the former as "containing" the predicate $\lambda y.$ not only y ($\lambda x.\ x$ loves y's wife), as also proposed by Evans (1977, pp. 95–97); whereas the latter has $\lambda y.$ not only y ($\lambda x.\ x$ loves Socrates's wife).

15 Richard's diagnosis was that (16a), despite its counter-intuitiveness in his scenario, is as true as (16b). But in later work he retracted this.

16 Pryor (2016) proposes that *graph-theoretic* constructions are a good way to abstract the shared core of these proposals. See also Strawson (1974/2004, pp. 45–46).

17 Compare Carnap's account of reports summarized at the start of §2 above. Salmon in fact worked with *guises* rather than Mentalese sentences; and he called the doxizing relation "BEL."

18 For example, Jacob (1991, pp. 93ff.).

19 See Fodor (1990a, esp. pp. 71–73), and Fitch (1984; 1986; 1987, ch. 4). Fitch (1996) defends a more specific view, of the sort associated below with Schiffer.

20 The following complaints are taken from Schiffer (1987, pp. 459–461; 1992; 1994, pp. 285–287; 1995, §5; 2000; 2003, pp. 39–46; 2008). Schiffer's own preferred theory is given in (1994; 2003, esp. ch. 2).

21 See Crimmins and Perry (1989), Crimmins (1992b; 1995b). Crimmins (1995c, pp. 200–201) objects to calling his view a "hidden indexical" theory.

22 Or what plays its role on their account: Crimmins thinks of T as a "thought map" describing a hypothetical attitude in terms of particular mental representations ("notions") the subject is assumed to have (1992b, chs 4–5).

23 In the hands of Stanley and Recanati, that phrase later acquired a more specific meaning than Perry understood by it (see Perry, 2001/2012, pp. 55–57; Korta and Perry, 2011, ch. 9; Crimmins, 1992b, pp. 16–17, 152).

24 The term "quasi-singular" comes from (Schiffer, 1978, p. 182). Recanati holds that in some cases complements don't express those but rather express content schemas (1993, ch. 18, n. 15 and §19.4), and in other cases express general propositions (§19.3).

Recanati's vocabulary and theoretical framework make his views somewhat hard to track. In (1993, ch. 3; 1995, §3) he distinguishes an utterance's "semantic" from its "truth-conditional" content. He understands the latter to be a Russellian proposition, and identifies it with "the proposition expressed" or "what is said" by a literal utterance (1993, pp. 54–55, 65). That is the notion that best fits our inquiry. (His "semantic" content belongs to more ambitious stories about how to explain successful understanding; see the discussion of Loar's stockbroker at Recanati, 1993, pp. 53ff.; 1995, pp. 184–185.) Recanati's view of unembedded sentences with "referential" terms is that their utterances have singular truth-conditional contents, as Russellians claim (though their semantic content will be richer). What I report in the text is that Recanati thinks *these terms* also have Russellian truth-conditional contents when embedded, but *the complement clauses that contain them*, as wholes, usually have richer, quasi-singular truth-conditional content (see 1993, pp. 395–397; 1995, pp. 188–190). In this regard, Recanati is closer to Schiffer's preferred (2003) view than to the HIT that Schiffer discusses more critically. Schiffer compares his version of HIT to Recanati's in Schiffer (2000).

25 Richard's theory is developed in (1990, chs 3–4), the central parts of which were also published as his (1989). The theory was then refined in his (1993; 1995; 1997, §§10–11; 2006; 2013b).

26 This quote is from Richard (2013b, p. 12). At this level of abstraction, Richard's view also resembles the neo-Fregean account of attitude reports offered in Forbes (1990). But I won't explore those connections.

27 So Richard's view has similarities and connections to the "ILF" theories advanced in Higginbotham (1986; 1991), Segal (1989), Larson and Ludlow (1993), Larson and Segal (1995, ch. 11), and Ludlow (2000).

28 Richard uses "accepts" in a technical sense: see his (1990, §1.5, and pp. 181–183, 211–213). He generalizes to attitudes other than belief in (ch. 3, n. 20).

29 Note that Richard never prohibited correlations from being many-one; he always allowed diverse expressions in a complement to correctly report beliefs that a subject in fact has by doxizing recurring Mentalese expressions. That is, if Anita accepts only (her Mentalese version of) (5), nonetheless (in some contexts) I can correctly say:

(1*) `Anita believes that Cicero admired Tully.`

That is, (in some contexts) (1*) can *allow* the reported thought to be cyclic. Richard doesn't think such a report would *attribute* a cyclic thought (1990, p. 217), but he left open whether it's *ever* possible for diverse terms in a complement to do so (pp. 214–217). In other words, he only endorsed the left-to-right direction of: recurring expressions in the complement → the report attributes a cyclic thought.

The cases that later caused Richard to partly withdraw that endorsement were given by Soames and Crimmins. We'll discuss them in §6 below.

30 Fine resists Putnam's specific proposal that sentences like (1) and (5) differ in "logical structure" (p. 41). But they are pursuing similar ideas.

31 My presentation isn't maximally explicit either: for conciseness, I help myself to some natural shorthands that I'll trust the reader to understand.

For further discussion of the "antinomy of the variable" and Fine's solution, see Kellenberg (2010) and Pickel and Rabern (2016). Compare the latter's solution to https://en.wikipedia.org/wiki/De_Bruijn_index (accessed 30 September, 2016), which is a more efficient technique commonly used in writing compilers.

32 One may have the concern that the natural extension of the given semantics to $\supset$ will conflate $[\![\forall x(Fx \supset \bot)]\!]$ with $[\![(\forall xFx) \supset \bot]\!]$. But this does not happen. Although both of those will depend on $[\![x, F, x, \bot]\!]$, they'll depend on it in different ways, and end up being non-equivalent as they should.

33 Soames (2010, p. 465) suggests that Fine should handle examples like `I asked Martha to call me` along the same lines as (5). At (pp. 41, 122–123), Fine considers handling anaphoric examples like (3d) in this way too. See Soames (2012, pp. 259ff.) for complaints about how this interacts with Fine's desire to have multiple readings of attitude reports.

34 King (1996; 1995) is paradigmatic of these. This tradition goes back to Lewis (1970) and Cresswell and von Stechow (1982), and has roots in Carnap's notion of "intensional isomorphism," discussed in §2 above. See also the later developments of King's view, and comparisons to other accounts of structured propositions, in King (2007; 2011).

As Fine remarks on his p. 54, he only needs to rely on enough structure to talk about the different appearances of some object in a proposition.

35 Fine also briefly addresses the possibility of handling empty names and "confused" names (see ch.2, nn. 4 and 14, pp. 126–127; and his 2010c).

36 See (pp. 54–57; also pp. 77–78). For other articulations of this idea, see Soames (1987b, p. 112, attributing it to unpublished work by Kaplan), Salmon (1986a, ch. 4, n. 4; 1992, pp. 59–60), and Pinillos (2011, pp. 319ff.). Higginbotham (1985, pp. 564ff.) uses a similar device, but his explanatory purposes are not the same.

An alternative to wires would be to tag all the leaves of a structured proposition with indices, understanding leaves to be linked when they have the same value and index.

Fine in fact defines his "coordination schemes" so as to allow wires spanning *multiple* structured propositions in a sequence.

37 On (p. 59), Fine might instead be read as saying that (1) expresses a coarser, Russellian proposition; but I will ignore that possibility because it ill fits the overall shape of his account. Further, this passage appears only a page after Fine introduces the idea of different "levels" of semantic value. (Recall we saw a similar idea in Richard, 1986–1987.) Our discussion here is focused on only the most specific level.

38 See King (2007, pp. 41–42, 218–222); he is following a suggestion made by Quine (1940/1981), and then repeated in Kaplan (1986, p. 244), Salmon (1986a, p. 156), and Soames (1989–1990, p. 204). See also Crimmins (1992b, ch. 4) and Evans (1977, pp. 88–96). (On p. 102, Evans permits the wires to cross sentence boundaries, and from pp. 104ff. they're also used to join "donkey pronouns" to their antecedents. In these cases the apparatus has to be interpreted differently.)

This appeal to wires shouldn't be *conflated with* the way they're appealed to in the work cited in note 36 above. It is a substantive question whether the present use can be a special case of the wires Fine is working with. Nor should either of these uses be conflated with the use of arrows in linguistic work on anaphora (for example, in Higginbotham, 1983; 1985; Moltmann, 2006, pp. 236ff.; Heim, 1998; McKay, 1991, pp. 724ff.; and the usage in Evans, 1977, from 102ff. mentioned above; see also Fiengo and May 1994's notion of "coindexing without c-command"). But as we'll see in §6 below, there are plausible connections between these.

39 That does seem to be Fine's thinking (see his p. 30 and n. 11, and pp. 115–117). See also Soames (2012, pp. 238–241). Recall also Richard's notion of "co-relativized" terms, mentioned in §3 above.

40 For the failure of necessity, see note 33 above. See also Fine's discussion of `Carl/Peter Hempel` at pp. 46–47. `Paderewski` cases might be counter-examples to sufficiency, though this turns on controversial issues. Compare Fine's example where one has stipulated the same definition for `glub` and `flox` but isn't explicitly aware of having done so (pp. 129–130). For a range of similar cases, see the works cited at Soames (2012, n. 5).

41 This quote is from (p. 40); see also (pp. 60ff.). As we'll see in the next section, Pinillos uses similar language when defining his notion of "*de jure* coreference." So too does Recanati: see his (2012, pp. 92, 106, 110). Schroeter uses similar language when defining the notion of expressions "*striking one* as being *de jure* coreferential": in her (2003, pp. 18ff.; 2007, pp. 600, 611; and 2008, pp. 110ff.), she talks of its being "obvious and rationally incontrovertible" to the subject that the expressions codesignate. In (2012) she adds the condition that this appearance be "epistemically basic." Unlike some of the others, Schroeter thinks the appearance of *de jure* coreference can be mistaken: see her (2007; 2008).

42 I say "in some way" because Fine rejects the common view that `believe` expresses a dyadic relation between a subject and the proposition expressed by its complement (see his discussion of (SB) at 2010b, pp. 476–478; also the move labeled F4 below).

43 See (pp. 93, 120–121, and his 2010b). Notably, what Fine says about the truth-conditions for the "strong *de dicto*" reading involves reference to the report itself, but he doesn't want to say that the *semantic content* of these reports is self-reflexive. The issue raised next in the text (and afterwards labeled F4) also poses difficulties.

Fine also posits a "pure *de re*" reading of attitude reports that ignores any coordination in its complement. See Soames (2012, appendix) for complaints about these ambiguities, and how they interact with the motivation and explanation Fine offers in the earlier parts of his book.

44 Though with multiple subjects, we confront special issues because Fine denies that interpersonal coordination will be transitive: see Fine's (pp. 98, 105ff.), also Taschek (1998, §5) and Richard (1990, pp. 210ff.; 1993, 129ff.). Crimmins (1992a, pp. 193–194; 1995a, pp. 387–390) criticizes this part of Richard's account. Taylor (2000, pp. 175ff.; 2003b, pp. 12ff.) also discusses such cases, but instead of abandoning transitivity, he denies that subjects are authoritative about whether they've introduced a new name.

45 As we saw in §3 above, Richard committed to the left-to-right direction of this. See also Pinillos's Principle of Attitude Closure, discussed below, also his (2011, p. 316); Taschek's "Logic Requirement" in his (1995, pp. 77, 81, and 86; 1998); and Soames's "Belief Coordination Principle" in his (2012). See also Taylor (2003b, §8) and Forbes (1987, pp. 24–25).

46 Compare the discussion of "transparency" in Recanati (2012, chs 10–11; and 2015, §I.iii, which replies to Ball, 2015); Schroeter (2007; 2008); and the discussion of "slow-switching" in Boghossian (1989; 1992; 1994; 2011; 2015) and many other sources surveyed in Parent (2013, §3.3). See also Salmon's discussion of recognizing the reappearance of an object in several propositions, in his (2012, pp. 423–424, 425–426, and 427ff.; 2015, pp. 455–456). Salmon argues that failures of such recognition aren't manifestations of linguistic incompetence or semantic ignorance.

We'll return to this issue with Pinillos, in the next section.

47 For discussion of Fine's work, and his replies, see Soames (2010), Lawlor (2010), Hovda (2010), and Fine (2010b; 2010c; 2010a). See also Soames (2012) and (1987b, whose relevance was mentioned at the end of §3 above); and Salmon (2012), Fine (2013), Salmon (2015). Further commentary includes Ostertag (2009), Rattan (2010), Sosa (2010), Bonardi (2013), Weiss (2014), and Pickel and Rabern (forthcoming).

48 See Salmon (1986a, ch. 4, n. 4), Ostertag (2009, p. 348), Sosa (2010, pp. 353, 356), Soames (1987b, pp. 112–113, 123; 2010, p. 467; 2012, pp. 237, 244–245), Salmon and Soames (1988, p. 13, n. 10), and Pinillos (2015, p. 326). Putnam (1954, p. 156) seems to embrace the semantic equivalence of (23a) and (23b).

Taschek acknowledges this problem for his own account, which anticipates Fine's in several ways: see his (1995, pp. 78–80; 1998, p. 327). He proposes a solution much like the "ambitious" one I go on to describe Fine giving.

49 See also Taschek (1995, pp. 81ff.; 1998, esp. §II).

50 Pinillos (2015, pp. 330–334) argues that if we attend to embeddings of (1)/(22) and (23a/b) in attitude reports, they will be "always accompanied, often implicitly," by other reports that are also asserted or (more likely) *presupposed* in the discourse. This is similar to the proposal I mentioned Richard offering. The view of reports that Pinillos ends up with is similar to the Crimmins and Perry view described in §3 above, albeit with presuppositions taking over some of the work of their "unarticulated constituents."

51 This term is from his (1992). In (1995), he defines a somewhat different notion of "Content," in terms of which (1) and (5) do count as different because they aren't inter-substitutable in attitude reports. This notion is more closely connected to Taschek's notion of "logical potential" than to his notion of "information content," which he stops using.

52 Pryor (2016) discusses connections between this work and the "multi-centered" views of content in Ninan (2012; 2013). See also Recanati (2012, §18.3; 2015, §I.iv).

53 See his (2011, p. 315; 2015, p. 325). Sometimes Pinillos gives examples using "slash names" like `Hesperus/Phosphorus`. Other times he uses anaphoric expressions like `there` or `that planet`. Pinillos's interpretation of all these cases is challenged by Goodsell (2014) and Contim (2016).

54 Compare Fine's quote about "sensibly raising the question" whether the things they designate are the same. Contim (2016) argues that this criterion of Pinillos's may be too liberal.

55 See Soames (1989–1990, example 15a; 1994, examples 8ab; 2010, example 26; 2012, example 34a). See also Higginbotham (1991, examples 38 and 45), Sider (1995, pp. 503–505), Crimmins (1992a, pp. 192, 195–196; 1995a, pp. 383–387, 391–392), and Cumming (2008, example KR).

56 See Soames (1994, example 34d; 2012, example 34b). See also Higginbotham (1991, examples 42 and 46).

57 A special twist on Soames's examples is when they're embedded inside other reports, for example `The children thought that Juan told Maria that he wasn't him`. See Soames (1989–1990, example 18; 1994, example 9). For the purpose of reporting the children's thinking about Juan, we'd want `he` and `him` to be *de jure* coreferential; but for the purpose of reporting the content of what Maria was allegedly told, we wouldn't.

Using the machinery from §7 below, the children's thought might be represented as: `((λ (he) (alias ([him he]) (disclaim-identity he Maria he (value him)))) Juan)`, where `disclaim-identity` represents its second operand as having been told something from which she could infer `Someone is not himself` iff the third and fourth operands are `aliased?`, as here they are not. Nonetheless, from the perspective of the embedding thought, `he` and `him` are `aliased?`.

58 In an earlier paper Higginbotham (1983, pp. 404–406) endorsed the idea that "referential linking" is transitive. But Higginbotham (1985, pp. 570–574) argues against this.

59 This is one of three main families of contemporary language that descend from Lisp. Scheme is a *family* of languages because it has many different "implementations" that extend or further specify the language in somewhat different ways. Details on how to run the example code supplied here inside some implementations of Scheme can be found at http://www.jimpryor.net/research/code/dejure.html (accessed September 30, 2016). The linebreaks, indentation, and variation between `()`s and `[]`s are all just stylistic choices.

60 I've assumed that in the evaluation of `(list (f 3) (f 3) (f 3))`, we evaluate the operands from left to right. If we proceeded in a different order, we'd get a different result. Different implementations of Scheme handle this differently.

61 What we will make use of are ideas that in other programming languages go by the keywords "same lvalue," and "call-" or "pass-by-reference." Everything exhibited below is expressible in (many implementations of) Scheme, using that language's ability to define *macros* that operate on program syntax before it gets evaluated. See the URL in note 59 above for details.

62 Thanks to Chris Barker, Simon Charlow, Cian Dorr, Kit Fine, Oliver Marshall, Kat Przyjemski, Daniel Rothschild, Nathan Salmon, Ken Shan, Seth Yalcin; audiences at Oxford, NYU/La Pietra, Bielefeld, Paris, Ohio State, UCLA, and Berkeley; an NYU seminar in fall 2012; and a conference there on Kit Fine in January 2013.

References

(When multiple versions are given, any page references are to the last.)

Asher, N. 1986. "Belief in discourse representation theory." *Journal of Philosophical Logic*, 15(2): 127–189.

Asher, N. 1987. "A typology for attitude verbs and their anaphoric properties." *Linguistics and Philosophy*, 10(2): 125–197.

Asher, N. 1989. "Belief, acceptance and belief reports." *Canadian Journal of Philosophy*, 19(3): 327–361.

Bach, K. 1987. *Thought and Reference*. Oxford: Clarendon.

Ball, D. 2015. "Indexicality, transparency and mental files." *Inquiry*, 58(4): 353–367.

Boghossian, P. 1989. "Content and self-knowledge." *Philosophical Topics*, 17(1): 5–26.

Boghossian, P. 1992. "Externalism and inference." *Philosophical Issues*, 2: 11–28.

Boghossian, P. 1994. "The transparency of mental content." *Philosophical Perspectives*, 8: 33–50.

Boghossian, P. 2011. "The transparency of mental content revisited." *Philosophical Studies*, 155(3): 457–465.

Boghossian, P. 2015. "Further thoughts on the transparency of mental content." In *Externalism, Self-Knowledge and Skepticism: New Essays*, edited by S. Goldberg, pp. 97–112. Cambridge: Cambridge University Press.

Bonardi, P. 2013. "Semantic relationism, belief reports, and contradiction." *Philosophical Studies*, 166(2): 273–284.

Campbell, J. 1987–1988. "Is sense transparent?" *Proceedings of the Aristotelian Society*, 88: 273–292.

Church, A. 1950. "On Carnap's analysis of statements of assertion and belief." *Analysis*, 10(5): 98–99. Reprinted in Linsky, 1971, pp. 168–170.

Church, A. 1954. "Intensional isomorphism and identity of belief." *Philosophical Studies*, 5(5): 65–73. Reprinted in Salmon and Soames, 1988, pp. 159–168.

Clapp, L. 2000. "Beyond sense and reference: an alternative response to the problem of opacity." In *The Pragmatics of Propositional Attitude Reports*, edited by K. M. Jaszczolt, pp. 43–75. Amsterdam: Elsevier.

Contim, F. D. V. 2016. "Mental files and non-transitive de jure coreference." *Review of Philosophy and Psychology*, 7(2): 365–388.

Cresswell, M., and A. von Stechow. 1982. "*De re* belief generalized." *Linguistics and Philosophy*, 5(4): 503–535.

Crimmins, M. 1992a. "Context in the attitudes." *Linguistics and Philosophy*, 15(2): 185–198.

Crimmins, M. 1992b. *Talk about Beliefs*. Cambridge, MA: MIT Press.

Crimmins, M. 1995a. "Contextuality, reflexivity, iteration, logic." *Philosophical Perspectives*, 9: 381–399.

Crimmins, M. 1995b. "Notional specificity." *Mind & Language*, 10(4): 464–477.

Crimmins, M. 1995c. "Quasi-singular propositions: the semantics of belief reports." *Aristotelian Society*, suppl. vol. 69: 195–209.

Crimmins, M., and J. Perry. 1989. "The prince and the phone booth: reporting puzzling beliefs." *Journal of Philosophy*, 86(12): 685–711.

Cumming, S. 2008. "Variabilism." *Philosophical Review*, 117(4): 525–554.

Dahl, O. 1973. "On so-called 'sloppy identity.'" *Synthese*, 26(1): 81–112.

Evans, G. 1977. "Pronouns, quantifiers, and relative clauses (I)." *Canadian Journal of Philosophy*, 7(3): 467–536. Reprinted in Evans, 1985, pp. 76–152.

Evans, G. 1980. "Pronouns." *Linguistic Inquiry*, 11(2): 337–362. Reprinted in Evans, 1985, pp. 214–248.

Evans, G. 1985. *Collected Papers*. Oxford: Clarendon.

Fiengo, R., and R. May. 1994. *Indices and Identity*. Cambridge, MA: MIT Press.

Fiengo, R., and R. May. 2006. *De Lingua Belief*. Cambridge, MA: MIT Press.

Fine, K. 2003. "The role of variables." *Journal of Philosophy*, 100: 605–610.

Fine, K. 2007. *Semantic Relationism*. Oxford: Blackwell.

Fine, K. 2010a. "Comments on Paul Hovda's 'Semantics as information about semantic values.'" *Philosophy and Phenomenological Research*, 81(2): 511–518.

Fine, K. 2010b. "Comments on Scott Soames' 'Coordination problems.'" *Philosophy and Phenomenological Research*, 81(2): 475–484.

Fine, K. 2010c. "Replies to Lawlor's 'Varieties of coreference.'" *Philosophy and Phenomenological Research*, 81(2): 496–501.

Fine, K. 2010d. "Semantic necessity." In *Modality*, edited by B. Hale and A. Hoffman, pp. 65–80. Oxford: Oxford University Press.

Fine, K. 2013. "Recurrence: a rejoinder." *Philosophical Studies*, 169(3): 1–4.

Fitch, G. 1984. "Two aspects of belief." *Philosophy and Phenomenological Research*, 45(1): 89–101.

Fitch, G. 1986. "Belief ascription." *Philosophical Studies*, 49(2): 271–280.

Fitch, G. 1987. *Naming and Believing*. Dordrecht, Netherlands: Reidel.

Fitch, G. 1996. "Representing beliefs." *Philosophy and Phenomenological Research*, 56(3): 597–609.

Fodor, J. 1990a. "Substitution arguments and the individuation of beliefs." In *Meaning and Method: Essays in Honor of Hilary Putnam*, edited by G. Boolos, pp. 63–78. Cambridge: Cambridge University Press. Reprinted in Fodor, 1990b, pp. 161–176.

Fodor, J. 1990b. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Forbes, G. 1987. "Indexicals and intensionality: a Fregean perspective." *Philosophical Review*, 96(1): 3–31.

Forbes, G. 1990. "The indispensability of *Sinn*." *Philosophical Review*, 99(4): 535–563.

Geach, P. 1962/1980. *Reference and Generality*, 3rd edn. Ithaca, NY: Cornell University Press.

Geach, P. 1965. "Logical procedures and the identity of expressions." *Ratio*, 7(2): 199–205. Reprinted in Geach, 1972, pp. 108–115.

Geach, P. 1972. *Logic Matters*. Berkeley, CA: University of California Press.
Geach, P. 1975a. "Back-reference." *Philosophia*, 5(3): 193–206.
Geach, P. 1975b. "Names and identity." In *Mind & Language*, edited by S. Guttenplan, pp. 139–158. Oxford: Clarendon.
Goodsell, T. 2013. "Mental files and their identity conditions." *Disputatio*, 5: 177–190.
Goodsell, T. 2014. "Is de jure coreference non-transitive?" *Philosophical Studies*, 167(2): 291–312.
Heim, I. 1998. "Anaphora and semantic interpretation: a reinterpretation of Reinhart's approach." In *The Interpretive Tract*, edited by U. Sauerland and O. Percus, pp. 205–246. Cambridge, MA: MIT Press.
Heim, I., and A. Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.
Higginbotham, J. 1983. "Logical form, binding, and nominals." *Linguistic Inquiry*, 14(3): 395–420.
Higginbotham, J. 1985. "On semantics." *Linguistic Inquiry*, 16(4): 547–593.
Higginbotham, J. 1986. "Linguistic theory and Davidson's program in semantics." In *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, edited by E. Lepore, pp. 29–48. Oxford: Blackwell.
Higginbotham, J. 1991. "Belief and logical form." *Mind & Language*, 6(4): 344–369.
Hovda, P. 2010. "Semantics as information about semantic values." *Philosophy and Phenomenological Research*, 81(2): 502–510.
Jacob, P. 1991. "Semantics and psychology: the semantics of belief-ascriptions." In *New Inquiries into Meaning and Truth*, edited by N. Cooper and P. Engel, pp. 83–109. New York: St Martin's Press / Harvester Wheatsheaf.
Kamp, H. 1984–1985. "Context, thought and communication." *Proceedings of the Aristotelian Society*, 85: 239–261.
Kamp, H. 1990. "Prolegomena to a structural account of belief and other attitudes." In *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*, edited by C. A. Anderson and J. Owens, pp. 27–90. Stanford: CSLI Press.
Kaplan, D. 1986. "Opacity." In *The Philosophy of W. V. Quine*, edited by L. E. Hahn and P. A. Schilpp, pp. 229–289. La Salle, IL: Open Court.
Kaplan, D. 1990. "Words." *Aristotelian Society*, suppl. vol. 64: 93–119.
Keenan, E. 1971. "Names, quantifiers, and the sloppy identity problem." *Papers in Linguistics*, 4(2): 211–232.
Kellenberg, A. 2010. "The antinomy of the variable." *Dialectica*, 64(2): 225–236.
King, J. 1995. "Structured propositions and complex predicates." *Noûs*, 29(4): 516–535.
King, J. 1996. "Structured propositions and sentence structure." *Journal of Philosophical Logic*, 25(5): 495–521.
King, J. 2007. *The Nature and Structure of Content*. Oxford: Clarendon.
King, J. 2011. "Structured propositions." *Stanford Encylopedia of Philosophy*. http://plato.stanford.edu/entries/propositions-structured/ (accessed September 29, 2016).
Korta, K., and J. Perry. 2011. *Critical Pragmatics: An Inquiry into Reference and Communication*. Cambridge: Cambridge University Press.
Larson, R., and P. Ludlow. 1993. "Interpreted logical forms." *Synthese*, 95(3): 305–355.
Larson, R., and G. Segal. 1995. *Knowledge of Meaning: An Introduction to Semantic Theory*. Cambridge, MA: MIT Press.
Lasnik, H. 1976. "Remarks on coreference." *Linguistic Analysis*, 2: 1–22. Reprinted in Lasnik, 1989, pp. 90–109.
Lasnik, H. 1989. *Essays on Anaphora*. Dordrecht, Netherlands: Kluwer.
Lawlor, K. 2002. "Memory, anaphora, and content preservation." *Philosophical Studies*, 109(2): 97–119.
Lawlor, K. 2010. "Varieties of coreference." *Philosophy and Phenomenological Research*, 81(2): 485–495.
Lewis, D. 1970. "General semantics." *Synthese*, 22(1): 18–67. Reprinted in Lewis, 1983, pp. 189–229.
Lewis, D. 1979. "Scorekeeping in a language game." *Journal of Philosophical Logic*, 8(1): 339–359. Reprinted in Lewis, 1983, pp. 233–249.

Lewis, D. 1983. *Philosophical Papers*, vol. 1. New York: Oxford University Press.
Linsky, L., ed. 1952/1970. *Semantics and the Philosophy of Language*. Champaign, IL: University of Illinois Press.
Linsky, L., ed. 1971. *Reference and Modality*. Oxford: Oxford University Press.
Ludlow, P. 1995. "Logical forms and the hidden indexical theory: a reply to Schiffer." *Journal of Philosophy*, 92(2): 102–107.
Ludlow, P. 1996. "The adicity of 'believes' and the hidden indexical theory." *Analysis*, 56(2): 97–101.
Ludlow, P. 2000. "Interpreted logical forms, belief attribution, and the dynamic lexicon." In *The Pragmatics of Propositional Attitude Reports*, edited by K. M. Jaszczolt, pp. 31–42. Amsterdam: Elsevier.
Mates, B. 1950. "Synonymity." In *Meaning and Interpretation*, edited by D. S. MacKay and G. P. Adams, pp. 210–226. Berkeley, CA: University of California Press. Reprinted in Linsky, 1952/1970, pp. 111–136.
McKay, T. 1991. "Representing *de re* beliefs." *Linguistics and Philosophy*, 14(6): 711–739.
Moltmann, F. 2006. "Unbound anaphoric pronouns: e-type, dynamic, and structured-propositions approaches." *Synthese*, 153(2): 199–260.
Nelson, M. 2005. "The problem of puzzling pairs." *Linguistics and Philosophy*, 28(3): 319–350.
Ninan, D. 2012. "Counterfactual attitudes and multi-centered worlds." *Semantics and Pragmatics*, 5(5): 1–57.
Ninan, D. 2013. "Self-location and other-location." *Philosophy and Phenomenological Research*, 87(2): 301–331.
Ostertag, G. 2009. "Review of Kit Fine, *Semantic Relationism*." *Australasian Journal of Philosophy*, 87(2): 345–349.
Parent, T. 2013. "Externalism and self-knowledge." *Stanford Encylopedia of Philosophy*. http://plato.stanford.edu/entries/self-knowledge-externalism/ (accessed September 29, 2016).
Partee, B. 1989. "Binding implicit variables in quantified contexts." In *Papers from CLS 25*, edited by C. Wiltshire, B. Music, and R. Graczyk, pp. 342–365. Chicago: Chicago Linguistic Society. Reprinted in Partee, 2004, pp. 259–281.
Partee, B. 2004. *Compositionality in Formal Semantics*. Oxford: Blackwell.
Perry, J. 1980. "A problem about continued belief." *Pacific Philosophical Quarterly*, 61(4): 317–332. Reprinted in Perry, 1993/2000, pp. 57–74.
Perry, J. 1993/2000. *The Problem of the Essential Indexical and Other Essays*, expanded edn. Oxford: Oxford University Press.
Perry, J. 2001/2012. *Reference and Reflexivity*, 2nd edn. Stanford: CSLI Press.
Pickel, B., and B. Rabern. 2016. "The antinomy of the variable: a Tarskian resolution." *Journal of Philosophy*, 113(3): 137–170.
Pickel, B., and B. Rabern. Forthcoming. "Does semantic relationism solve Frege's puzzle?" *Journal of Philosophical Logic*. DOI: 10.1007/s10992-016-9420-z. http://www.semanticsarchive.net/Archive/GQ3ODk2M/relationism_Frege1.pdf (accessed September 29, 2016).
Pinillos, N. A. 2011. "Coreference and meaning." *Philosophical Studies*, 154(2): 301–324.
Pinillos, N. A. 2015. "Millianism, relationism, and attitude ascriptions." In *On Reference*, edited by A. Bianchi, pp. 322–334. Oxford: Oxford University Press.
Pryor, J. 2016. "Mental graphs." *Review of Philosophy and Psychology*, 7(2): 309–341. http://www.jimpryor.net/research/papers/Graphs.pdf (accessed September 29, 2016).
Putnam, H. 1954. "Synonymity, and the analysis of belief sentences." *Analysis*, 14(5): 114–122. Reprinted in Salmon and Soames 1988, pp. 149–158.
Quine, W. V. O. 1940/1981. *Mathematical Logic*, rev. edn. Cambridge, MA: Harvard University Press.
Rattan, G. 2010. "*Semantic Relationism*, by Kit Fine." *Mind*, 118(472): 1124–1131.
Recanati, F. 1993. *Direct Reference: From Language to Thought*. Oxford: Blackwell.
Recanati, F. 1995. "Quasi-singular propositions: the semantics of belief reports." *Aristotelian Society*, suppl. vol. 69: 175–193.

Recanati, F. 2005. "Deixis and anaphora." In *Semantics vs. Pragmatics*, edited by Z. G. Szabó, pp. 286–316. Oxford: Oxford University Press.
Recanati, F. 2012. *Mental Files*. Oxford: Oxford University Press.
Recanati, F. 2013. "Mental files: replies to my critics." *Disputatio*, 5(36): 207–242.
Recanati, F. 2015. "Replies." *Inquiry*, 58(4): 408–437.
Reinhart, T. 1983. *Anaphora and Semantic Interpretation*. London: Croom Helm.
Richard, M. 1983. "Direct reference and ascriptions of belief." *Journal of Philosophical Logic*, 12(4): 425–452. Reprinted in Richard, 2013a, pp. 26–47.
Richard, M. 1986–1987. "Attitude ascriptions, semantic theory, and pragmatic evidence." *Proceedings of the Aristotelian Society*, 87: 243–262. Reprinted in Richard, 2013a, pp. 65–79.
Richard, M. 1987. "Quantification and Leibniz's Law." *Philosophical Review*, 96(4): 555–578. Reprinted in Richard, 2013a, pp. 48–64.
Richard, M. 1989. "How I say what you think." *Midwest Studies in Philosophy*, 14(1): 317–337. Reprinted in Richard, 2013a, pp. 80–99.
Richard, M. 1990. *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*. Cambridge: Cambridge University Press.
Richard, M. 1993. "Attitudes in context." *Linguistics and Philosophy*, 16(2): 123–148. Reprinted in Richard, 2013a, pp. 100–120.
Richard, M. 1995. "Defective contexts, accommodation, and normalization." *Canadian Journal of Philosophy*, 25(4): 551–570. Reprinted in Richard, 2013a, pp. 121–136.
Richard, M. 1997. "What does commonsense psychology tell us about meaning?" *Noûs*, 31(1): 87–114.
Richard, M. 2006. "Meaning and attitude ascriptions." *Philosophical Studies*, 128(3): 683–709. Reprinted in Richard, 2013a, pp. 246–262.
Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.
Richard, M. 2013a. *Context and the Attitudes*. Oxford: Oxford University Press.
Richard, M. 2013b. "Introduction." In *Context and the Attitudes*, pp. 1–25. Oxford: Oxford University Press.
Richard, M. 2015. *Truth and Truth Bearers*. Oxford: Oxford University Press.
Salmon, N. 1986a. *Frege's Puzzle*. Cambridge, MA: MIT Press.
Salmon, N. 1986b. "Reflexivity." *Notre Dame Journal of Formal Logic*, 27(3): 401–429. Reprinted in Salmon, 2007, pp. 32–57.
Salmon, N. 1989. "Illogical belief." *Philosophical Perspectives*, 3: 243–285. Reprinted in Salmon, 2007, pp. 193–223.
Salmon, N. 1992. "Reflections on reflexivity." *Linguistics and Philosophy*, 15(1): 53–63. Reprinted in Salmon, 2007, pp. 58–66.
Salmon, N. 2001. "The very possibility of language: a sermon on the consequences of missing church." In *Logic, Meaning, and Computation: Essays in Memory of Alonzo Church*, edited by C. A. Anderson and M. Zelëny, pp. 573–595. Dordrecht, Netherlands: Kluwer.
Salmon, N. 2005. *Metaphysics, Mathematics, and Meaning*. Oxford: Clarendon.
Salmon, N. 2006a. "Pronouns as variables." *Philosophy and Phenomenological Research*, 72(3): 656–664. Reprinted [sic] in Salmon, 2005, pp. 399–406.
Salmon, N. 2006b. "The resilience of illogical belief." *Noûs*, 40(2): 369–375. Reprinted in Salmon, 2007, pp. 224–229.
Salmon, N. 2007. *Content, Cognition, and Communication*. Oxford: Clarendon.
Salmon, N. 2010. "Lambda in sentences with designators: an ode to complex predication." *Journal of Philosophy*, 107(9): 445–468.
Salmon, N. 2012. "Recurrence." *Philosophical Studies*, 159(3): 407–441.
Salmon, N. 2015. "Recurrence again." *Philosophical Studies*, 172(2): 445–457.
Salmon, N., and S. Soames, eds. 1988. *Propositions and Attitudes*. Oxford: Oxford University Press.
Saul, J. 1993. "Still an attitude problem." *Linguistics and Philosophy*, 16(4): 423–435.

Schiffer, S. 1977. "Naming and knowing." *Midwest Studies in Philosophy*, 2(1): 28–41.
Schiffer, S. 1978. "The basis of reference." *Erkenntnis*, 13(1): 171–206.
Schiffer, S. 1987. "The 'Fido'-Fido theory of belief." *Philosophical Perspectives*, 1: 455–480.
Schiffer, S. 1992. "Belief ascription." *Journal of Philosophy*, 89(10): 499–521.
Schiffer, S. 1994. "A paradox of meaning." *Noûs*, 28(3): 279–324.
Schiffer, S. 1995. "Descriptions, indexicals, and belief reports: some dilemmas (but not the ones you expect)." *Mind*, 104(413): 107–131.
Schiffer, S. 1996. "The hidden indexical theory's logical-form problem: a rejoinder." *Analysis*, 56(2): 92–97.
Schiffer, S. 2000. "Propositional attitudes in direct-reference semantics." In *The Pragmatics of Propositional Attitude Reports*, edited by K. M. Jaszczolt, pp. 13–30. Amsterdam: Elsevier.
Schiffer, S. 2003. *The Things We Mean*. Oxford: Oxford University Press.
Schiffer, S. 2008. "Propositional content." In *Oxford Handbook of Philosophy of Language*, edited by E. Lepore and B. Smith. Oxford: Oxford University Press.
Schroeter, L. 2003. "Gruesome diagonals." *Philosophers' Imprint*, 3(3): 1–23.
Schroeter, L. 2007. "The illusion of transparency." *Australasian Journal of Philosophy*, 85(4): 597–618.
Schroeter, L. 2008. "Why be an anti-individualist?" *Philosophy and Phenomenological Research*, 77(1): 105–141.
Schroeter, L. 2012. "Bootstrapping our way to samesaying." *Synthese*, 189(1): 177–197.
Segal, G. 1989. "A preference for sense and reference." *Journal of Philosophy*, 86(2): 73–89.
Sider, T. 1995. "Three problems for Richard's theory of belief ascription." *Canadian Journal of Philosophy*, 25(4): 487–513.
Soames, S. 1985. "Lost innocence." *Linguistics and Philosophy*, 8(1): 59–71.
Soames, S. 1987a. "Direct reference, propositional attitudes, and semantic content." *Philosophical Topics*, 15(1): 47–87. Reprinted in Soames, 2009b, pp. 33–71.
Soames, S. 1987b. "Substitutivity." In *On Being and Saying: Essays for Richard Cartwright*, edited by J. J. Thomson, pp. 99–132. Cambridge, MA: MIT Press.
Soames, S. 1989–1990. "Pronouns and propositional attitudes." *Proceedings of the Aristotelian Society*, 90: 191–212.
Soames, S. 1994. "Attitudes and anaphora." *Philosophical Perspectives*, 8, 251–272. Reprinted in Soames, 2009b, pp. 111–135.
Soames, S. 1995. "Beyond singular propositions?" *Canadian Journal of Philosophy*, 25(4): 515–549.
Soames, S. 2002. *Beyond Rigidity*. Oxford: Oxford University Press.
Soames, S. 2005. "Naming and asserting." In *Semantics vs. Pragmatics*, edited by Z. G. Szabó, pp. 356–382. Oxford: Oxford University Press. Reprinted in Soames, 2009a, pp. 251–277.
Soames, S. 2009a. *Philosophical Essays*, vol. 1. Princeton, NJ: Princeton University Press.
Soames, S. 2009b. *Philosophical Essays*, vol. 2. Princeton, NJ: Princeton University Press.
Soames, S. 2010. "Coordination problems." *Philosophy and Phenomenological Research*, 81(2): 464–474.
Soames, S. 2012. "Two versions of Millianism." In *Reference and Referring*, edited by W. Kabasenche, M. O'Rourke, and M. Slater, pp. 83–118. Cambridge, MA: MIT Press. Reprinted in Soames, 2014, pp. 231–264.
Soames, S. 2014. *Analytic Philosophy in America and Other Historical and Contemporary Essays*. Princeton, NJ: Princeton University Press.
Sosa, D. 2010. "The Fine line." *Analysis*, 70(2): 347–358.
Strawson, P. 1974/2004. *Subject and Predicate in Logic and Grammar*, rev. edn. London: Methuen.
Taschek, W. 1992. "Frege's puzzle, sense, and information content." *Mind*, 101(404): 767–791.
Taschek, W. 1995. "Belief, substitution, and logical structure." *Noûs*, 29(1): 71–95.
Taschek, W. 1998. "On ascribing beliefs: content in context." *Journal of Philosophy*, 95(7): 323–353.
Taylor, K. 2000. "Emptiness without compromise." In *Empty Names, Fiction, and the Puzzles of Non-Existence*, edited by A. Everett and T. Hofweber, pp. 17–36. Stanford: CSLI Press. Reprinted in Taylor, 2003a, pp. 167–190.

Taylor, K. 2003a. *Reference and the Rational Mind*. Stanford: CSLI Press.
Taylor, K. 2003b. "What's in a name?" In *Reference and the Rational Mind*, pp. 1–32. Stanford: CSLI Press.
Taylor, K. 2015. "Names as devices of explicit co-reference." *Erkenntnis*, 80(2): 235–262.
Tomioka, S. 1999. "A sloppy identity puzzle." *Natural Language Semantics*, 7(2): 217–241.
Weiss, M. 2014. "A closer look at manifest consequence." *Journal of Philosophical Logic*, 43(2): 471–498.
Wiggins, D. 1976a. "Frege's problem of the morning star and the evening star." In *Studies on Frege II. Logic and the Philosophy of Language*, edited by M. Schirn, pp. 221–255. Stuttgart: Bad Canstatt.
Wiggins, D. 1976b. "Identity, necessity, and physicalism." In *Philosophy of Logic: Papers and Discussions*, edited by S. Körner, pp. 96–132, 159–182. Berkeley, CA: University of California Press.
Williams, E. 1977. "Discourse and logical form." *Linguistic Inquiry*, 8(1): 101–139.

Glossary

Absolute versus relative identity: The contrast between absolute and relative identity has its home in the debate over Peter Geach's Relative Identity Thesis, given in "Identity," *Review of Metaphysics*, 21 (1967). An absolute-identity relation is an equivalence relation which satisfies Leibniz's Law; a relative-identity relation is an equivalence relation which does not. It is uncontroversial that some relations are mere relative-identity relations. In fact, of course, this is true of all equivalence relations not satisfying Leibniz's Law (q.v.). What is controversial is whether an equivalence relation expressible in the form '*x* is the same A as *y*,' where 'A' is a sortal term, can be a mere relative-identity relation, and whether absolute identity is expressible at all. Geach answers the first of these questions affirmatively and the second negatively.

Absolute versus relative notions of necessity/possibility: When 'necessity' and 'possibility' are qualified by prefixing 'physical', 'natural', 'biological' or the like, the notions expressed are probably best understood as relative to the assumption of the laws of the discipline to which the adjective alludes. To say, for example, that it is physically necessary that *P* is to claim that, given the laws of physics, it must be true that *P* (i.e., that it is a logical consequence of the laws of physics that *P*). Similarly, the claim that it is physically possible that *P* would normally be understood as the claim that the laws of physics do not exclude its being true that *P* (i.e., that it is logically consistent with the laws of physics that *P*). If, as is plausible, we allow that the universe might have behaved according to different physical laws, so that what is physically necessary, as things are, might not have been true, our notion of physical necessity could be said to be merely relative. Sometimes, however, when we claim that it is necessary that *P*, we mean to deny that there is any possibility of things being otherwise – we are claiming absolute necessity. Such an absolute notion is probably intended in claims about logical or metaphysical necessity. In terms of possible worlds, it is absolutely necessary that *P* if *P* holds true at all possible worlds without qualification; whereas it suffices for it to be physically necessary

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

that *P* that *P* holds true at all physically possible worlds (i.e., all possible worlds where the actual physical laws hold), and similarly for other relative kinds of necessity.

Abstract objects: The distinction between abstract and concrete entities is usually thought to be exhaustive and mutually exclusive. A popular view is that the hallmark of the concrete is existence in physical space and/or time. As a corollary of this, it is often held that abstract entities lack causal powers and are consequently incapable of entering into causal relations with other things (though this threatens to make our knowledge of them problematic). In opposition to this way of characterizing the abstract/concrete divide, others have been proposed: for instance, some philosophers characterize abstract entities as ones which depend logically for their existence upon the existence of certain other entities (as, for example, a smile is said to depend for its existence upon the face whose smile it is), while others characterize abstract entities as the products of some sort of mental or logical process of 'abstraction' from concepts (as when numbers and geometrical shapes are said to be such products).

Acquisition argument: Common term for an objection (associated particularly with Michael Dummett) to any semantic theory which allows the nature of speakers' private states (see **Privacy**) to influence the meaning of the words they use: it cannot explain how a learner ever acquires an understanding of the language. The argument turns on the point that the meanings to be learned would depend on the nature of the private states of the already-competent speaker, something which cannot, by definition, be known to anyone else. The Acquisition Argument is a close relation of the arguments from Communicability (q.v.) and Manifestation (q.v.). (See Chapter 11, MEANING AND PRIVACY.) Interest in these arguments has been aroused by Dummett's suggestion that they support an 'anti-realist' approach to semantics, shifting emphasis away from the conditions under which a sentence is true, perhaps Evidence-transcendently (q.v.), towards those under which a speaker might properly assert it. (See also Chapter 20, REALISM AND ITS OPPOSITIONS.)

Analyses, reductive and reciprocal: An analysis of a concept breaks up a given concept (the analysandum concept) into its component concepts (the analysans concepts). An analysis is represented by a biconditional, thus (in the case of meaning): X means that p iff … [favored analysans concepts]. Analyses are usually taken to be reductive in nature, which is to say that the analysans concepts are held to be more fundamental or basic (in some sense to be specified) than the analysandum concept. But analyses may also be reciprocal in nature, meaning that the concepts on either side of the biconditional are seen as on a par – the idea being that whilst the proposed analysis is non-reductive, it illuminates the concepts involved by drawing out important links between them.

Asymmetric Dependence Theory (ADT): ADT is a development of the intuitive idea that the truth-conditions of a thought are resilient with respect to other causes of the thought. According to ADT, a Mentalese (q.v.) predicate C refers to the property P if C locks onto P (J. Fodor, *Psychosemantics*, MIT Press, 1987; *A Theory of Content*, MIT Press, 1990). C locks onto P just in case (i) it is a law that Ps cause Cs, (ii) there is some Q other than P such that Qs cause Cs, and for any Q distinct from P, if Qs cause Cs then the causal connection between Qs and Cs depends on the Ps-cause-Cs law, but not the other way round. In other words, if Qs failed to cause Cs it would still be the case that Ps caused Cs; but if Ps failed to cause Cs then Qs wouldn't cause them either. ADT solves the Disjunction problem (q.v.), since not all the causes of C constitute its reference. However, the theory is difficult to evaluate, since it is not clear exactly what the dependence relation between laws (or causal relations) is, and whether it is naturalistic.

Binding: In semantic and syntactic theory, 'binding' refers to the relation between an anaphoric item (pronouns, reciprocals, and other pro-forms) and its antecedent such that the former is referentially dependent on the latter. Modern inquiry into binding goes back at least to the 1960s, but is often summarized by so-called *classic binding theory* (CBT) formulated in the early 1980s by Noam Chomsky. CBT contains three principles:

(A) Reflexives (*herself*, *themselves*, etc.) are bound by a local antecedent.
(B) Pronouns (when bound) are bound by a non-local antecedent.
(C) Referential terms (e.g., names, bare plurals, etc.) are not bound (free).

The precise characterization of locality is a long-standing matter of dispute, but it is traditionally understood to be a syntactic relation of hierarchical distance relative to certain categories (intuitively, one might think of *local* as meaning *at least local to the clause*; so, (A) says that a reflexive must be bound by an item within the same clause as it). One item (a binder) binds another (the bindee), if they share an index and are in the right syntactic configuration in adherence to the stated principles. The content of the three principles are exemplified below:

(1) a Sam_i thought Dave_j liked $\text{himself}_{*i/j}$ [Only *Dave* is local enough to bind *himself*]
b Sam_i thought Dave_j liked $\text{him}_{i/*j}$ [Only *Sam* is non-local enough to bind *him*]
c He_i thought she_j liked $\text{Jo}_{*i/*j}$ [Referential *Jo* lacks any antecedent]

The linguistic status of binding is currently unsettled; whether, that is, it is a syntactic, semantic, or pragmatic phenomenon, or some admixture of the three. Many of these uncertainties go back to the 1970s, and gain a new urgency by various scruples against the free use of indexes within the so-called *minimalist program*, as well as the development of alternative syntactic accounts that reject the fundamental role of configurational or merely syntactic conditions on binding.

Whatever the ultimate fate of binding, the notion is not to be confused with co-reference, which is an extra-linguistic relation. For example, (2) is acceptable on the indicated reading, but is a Principle C violation:

(2) Only Bill_i likes Bill_i

(2) does not so much refute Principle C as indicate an obviation of the principle; after all, that the two tokens of *Bill* in (2) refer to the one individual is not a matter of linguistic competence. Moreover, binding applies to bound-variable anaphora, involving quantifier phrases, where the antecedent is not referential, and so the relation of binding cannot be one of mere co-reference. Thus, if co-referential items are inter-substitutable *salva veritate* (in the relevant contexts), then substituting referential terms for pronouns will result in (at worst) Principle C violations:

(3) a Bill_i thought he_i was a genius
b Bill_i thought Bill_i was a genius

Such substitution produces a change of meaning (truth-conditions) where the binding is bound-variable anaphora:

(4) a $\text{Every philosopher}_i$ thinks he_i is a genius
b $\text{Every philosopher}_i$ thinks every philosopher_i is a genius

CBT is designed to cover both overt and covert items within a syntactic structure. Thus, binding is appealed to as a diagnostic for covert items. For example, according

to Principle A, a reflexive requires a local binder, but a reflexive may occur in the absence of any such *overt* binder:

(5) Mary wanted to leave by herself/*himself

The difference in the indicated judgments in (5) suggests that the non-finite clause has a covert subject (commonly designated PRO) that is both referentially dependent on *Mary* and binds *herself*, but cannot bind *himself* without an agreement error arising. If (5) (apparently) occurs as a complement of an embedded question, no potential for an agreement error arises:

(6) Sam wondered who Mary wanted to leave by herself/himself

This is explained by the covert subject of the non-finite clause being a covert trace or copy of *who*, which, being unmarked for gender, may bind either reflexive. In effect, therefore, (5) does not occur in (6) as a syntactic constituent, for the respective underlying forms diverge with respect to their constituent covert items.

That binding holds across both overt and covert items has been held to militate for certain syntactic/semantic accounts of philosophically contentious constructions, such as quantifier domain restriction and predicates of personal taste.

Bivalence: The principle of bivalence asserts that every statement is true or false. It should be distinguished from the Law of Excluded Middle, according to which every instance of the schema 'P or not-P' is true. Under the assumption that any false statement's negation is true, bivalence entails Excluded Middle. But Excluded Middle does not entail bivalence, since many-valued and supervaluational semantics may validate the former but do not validate the latter (see **Law of Excluded Middle**). Unrestricted endorsement of bivalence has been taken, especially by Michael Dummett, as the hallmark of a realist position with regard to statements of some given kind. On some views about vagueness, bivalence fails for vague statements, which are held to be neither true nor false in borderline cases.

Causal theory of reference: A causal theory of reference is a theory that attempts to explain the nature of the relation between the use of a referring expression and the referent – to say what it is about the use of an expression in virtue of which it has the referent that it has. What all such theories have in common is the thesis that the reference relation should be explained in terms of a causal or explanatory connection between the object that is the referent and the speaker's use of the expression: the referent of a name, as used on a particular occasion, is that object which plays a certain role in the causal process that results in the speaker's use of the name on that occasion. The constructive task of such a theory is to say more specifically how an object must be causally related to the use of the expression in order to be the referent of the expression.

Character, meaning, and content: The terms 'meaning' and 'content' are used in many ways in philosophy, and it is always important to check what a particular author may have in mind. In the usage of Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS, meanings are properties of types of expressions, fixed by the rules of language; contents are properties of specific utterances. In the case of indexicals and expressions containing indexicals, the content of an utterance will not be fully determined by the meaning of the expressions used, but will also depend on context. "Character" is often used instead of "meaning," following David Kaplan in his work on demonstratives.

Cognitive command: The cognitive-command constraint is proposed by Crispin Wright as one of several features or marks in virtue of which realism concerning the subject-matter of a given discourse whose characteristic statements qualify for at least Minimal truth (q.v.) may be maintained. Roughly, a discourse satisfies this constraint just in case it is *a priori* that differences of opinion arising within it can be satisfactorily explained only in terms of something worth describing as a cognitive shortcoming in one or other of the disagreed parties – at least one of the parties to the disagreement is lacking relevant information, or her assessment of the data is distorted by prejudice or idiosyncratic standards, or some such. The intended contrast is with cases in which disagreement may persist after all the relevant information is in, say because the disagreed parties diverge in their affective (non-cognitive) reactions to the facts – as is plausibly the case with divergent judgments about what is beautiful or funny, for example. (See also **Euthyphro contrast, Wide cosmological role**.)

Coherence: Coherence theories of truth hold that the truth of a belief consists in its coherence with the main body of our beliefs; coherence theories of knowledge hold that a belief's justification (but not necessarily its truth) consists in the same coherence. Those who hold these theories vary over what they understand by "coherence." Usually it is agreed that a coherent system must be consistent, and the differences are over what further requirements there are; though clearly they must be determined by the main body of our beliefs. Breadth, together with some sort of overall simplicity, are often required. What these requirements amount to depends on the role assigned to experience. Some hold that the coherent system must fit (in some manner hard to specify) with the content of experience as presented preconceptually. Some hold that experience can bear on the coherent system only by providing us with experiential beliefs, but that these beliefs have a special status: their truth (or their justification) is direct, not a matter of coherence, though to all other beliefs the coherence theory applies. Some, again, hold that their truth (or their justification) does consist in coherence: perhaps the most consistent line for a coherence theorist.

Communicability argument: Common term for an objection (associated particularly with Michael Dummett) to any semantic theory which allows the nature of speakers' private states (see **Privacy**) to influence the meaning of the words they use: it cannot explain how there can be communication between speaker and hearer. For if the former's meaning depends on the nature of their private states, what they mean is (by the definition of a private state, see **Privacy**) unknowable to the hearer. Yet for communication to take place the hearer must know what the speaker means. The Communicability Argument is a close relation of the arguments from Acquisition (q.v.) and Manifestation (q.v.). (See Chapter 11, MEANING AND PRIVACY.) Interest in these arguments has been aroused by Dummett's suggestion that they support an 'anti-realist' approach to semantics, shifting emphasis away from the conditions under which a sentence is true – perhaps Evidence-transcendently (q.v.) – towards those under which a speaker might properly assert it.

Compositionality: A theory of meaning is said to be compositional when (a) it has only finitely many axioms and (b) it delivers up meaning-specifying theorems on the basis of those axioms in such a way that the semantic structure of the sentences of the language is thereby exhibited. Intuitively, a compositional theory of meaning would serve up a meaning-specifying theorem for "The man with the red nose is drinking whisky" on the basis of the meanings of its constituent words and their mode of syntactic combination. An example of a non-compositional theory of meaning for a language with only finitely

many sentences would be a long list of meaning-specifying theorems, one for each sentence of the language; an example of a non-compositional theory of meaning for a language with infinitely many sentences would be provided by an infinitary axiom schema such as *A is True iff P*, where "*P*" could be replaced by any declarative sentence of the language concerned and "*A*" by the quotational name of that sentence (assuming, for simplicity, that the correct form of a meaning-specification is a statement of truth-conditions). The construction of compositional theories of meaning is thought by some philosophers to throw light on such phenomena as Semantic creativity and the learnability (q.v.) of natural languages. (See Chapter 12, TACIT KNOWLEDGE.)

Context: The context of an utterance is the situation in which it occurs. The context is often needed to resolve questions about what words stand for. Context can be relevant, however, in different ways. Sometimes it is relevant to determining which expressions are used. Sometimes it is relevant to the resolution of ambiguities: to determine with which meanings expressions have been used. In the case of indexicals, context remains relevant when both the identity and the meaning of expressions are known, for the meanings of the expressions are rules that fix the designation of the expressions relative to contextual factors.

Context Principle: A principle enunciated in several forms by Frege in his *Foundations of Arithmetic* (1884: Eng. trans. 1959), asserting that it is only in the context of a complete sentence or proposition that a word has meaning (*Bedeutung*). In *Foundations*, Frege undoubtedly appeals to the principle to justify his view that it is not necessary for a word to have meaning that we should be able to point to what it stands for, and thence in arguing against psychologistic tendencies which seek to identify the meaning of number-words with mental entities. It is also taken by him, at least during his middle period, as justifying the procedure of defining terms contextually. Both its interpretation and its wider bearing upon issues in the philosophy of language and ontology are controversial. This is in part because, at the time of writing *Foundations*, Frege had not explicitly drawn his celebrated distinction between sense (*Sinn*) and reference (*Bedeutung*), so that it is unclear whether the principle should be taken as applying to sense, or to reference, or to both. Taken as applying to sense, the principle may be seen as underpinning his view that the sense of sub-sentential expressions consists in their contribution to the sense of complete sentences containing them, and so as supporting a position intermediate between semantic atomism at one extreme and Semantic holism (q.v.) at the other. Taken as applying to reference, the principle may be seen as underpinning the view (to which Frege also subscribes) that it suffices for terms to have reference that they figure as parts of more complex expressions (e.g., sentences) which have reference.

Convention: The fundamental idea in the application of the notion of convention to language is to capture the arbitrary nature of the association between a word and its meaning. This *idea* is not in dispute by philosophers, despite the fact that at least one philosopher does deny that language is conventional. To understand this it is necessary to understand that the concept of convention has come to be associated with the idea of a kind of rational control by speakers over the meaning of their words. It is this that some philosophers dispute. Thus one finds two positions in the literature: (1) language is conventional in the sense of being both arbitrary and under speakers' rational control, and (2) language is conventional only in the sense of being arbitrary. At least one philosopher (Davidson) claims to reject the idea of language as governed by conventions altogether (see postscript to Chapter 13, RADICAL INTERPRETATION); however, on closer inspection

one finds that the idea that meaning is arbitrary is still retained (only the association between this idea and convention is abandoned).

Convention T: see **Criterion of material adequacy**

Counterpart theory: An alternative semantics for quantified modal logic, due to David Lewis. In standard semantics, a sentence '*Possibly, Fa*' is true iff there is a possible world in which *a* is *F*. In counterpart-theoretic semantics, '*Possibly, Fa*' is true iff there is a possible world in which some counterpart of *a* is *F*. In Lewis's own version the counterpart relation is determined by similarity, but this philosophical view about the nature of counterparthood is only one of many that are consistent with the semantics.

Criterion of identity: The general notion of a criterion of identity is the notion of a standard by which identity is to be judged. A paradigm is Frege's example: the criterion of identity for directions is parallelism of lines. In this case the criterion of identity for one type of object is a relation between objects of another type. One question about criteria of identity is whether this must always be the case, or whether a criterion of identity for one type of object can be a relation between objects of the same type. A second question is whether a criterion of identity needs to be a relation at all. A standard, though controversial view is that reference to an object is only possible against the background of a criterion of identity, and hence that any proper name must have as part of its sense a criterion of identity.

Criterion of material adequacy: When Tarski set out to define "true sentence of language L," by the systematic determination of its extension, he stipulated that the definition should be formally rigorous, should make no use of semantic notions, and should in addition be *materially adequate*, or, in other words, should grasp the current meaning of the notion of truth as it is known intuitively. The definition must be faithful to the substance of the notion, in other words. What this involved (see Chapter 2, MEANING AND TRUTH-CONDITIONS, §§16–17) was that the definition should *entail*, for each sentence of the language for which "true sentence" is to be defined, a biconditional of the form:

True x if and only if p

where the "x" holds a place for a designation of the sentence in question, and "p" holds a place for a translation of the sentence into the metalanguage in which the definition is being constructed. In the case where the object-language is a proper part of the metalanguage, the object-language sentence will count as a translation of itself. The same sentence, say "snow is white," is then referred to on the left-hand side and used on the right-hand side of the biconditional that the truth-definition must entail.

NB. A biconditional such as '"snow is white" is true if and only if snow is white' would be called in propositional logic a material equivalence. This use of the word "material" in logic has absolutely nothing to do with what Tarski means by material adequacy in his statement of Convention T. The material adequacy of a truth-definition involves the fidelity of the definition to the intuitive notion of truth. Material adequacy is seen by Tarski as a *substantial* requirement.

***De dicto* and *de re* ascriptions:** One can ascribe a property – or a "mode" of having a property, such as having it necessarily – to what a sentence says (a *dictum*) or to an individual (a *res*). Ascriptions of the first sort are *de dicto*, of the latter *de re*. To say that the claim that my mother had no children is possibly true is to make a (silly) *de dicto* ascription; to say, of my mother, that she might have been childless, is to make a (true) *de re* one. *De re*, but not *de dicto*, modal ascriptions have implications about the essential properties of an

object. Sentences with quantifier phrases and operators (such as 'my mother might not have had a child') are often ambiguous between a *de dicto* and a *de re* reading. Some, such as Quine, have challenged the intelligibility of *de re* modal claims and of devices, such as quantification into modal contexts, typically used to make them.

De dicto and *de re* propositional-attitude ascriptions are distinguished as above: the first relates one to a *dictum* (a proposition); the latter to an individual and an attribute. In a *de re* ascription, such as 'Kristen believes, of my mother, that she is childless,' the position which picks out the *res* (here, that of 'my mother') is transparent; in a *de dicto* reading of 'Kristen believes that my mother is childless,' the position of 'my mother' is opaque. It is a matter of dispute whether this distinction, between ways of *talking* about attitudes, corresponds to a distinction between *kinds* of attitudes – whether, for example, there is a special kind of *de re* belief which involves an epistemically significant connection or acquaintance with an object.

***De re* senses:** A *de re* sense is a mode of presentation of an object which cannot be entertained if that object does not exist. *De re* senses were first introduced as such in the works of John McDowell, such as "On the sense and reference of a proper name," *Mind*, 86 (1977), pp. 159–185, and Gareth Evans, as in "Understanding demonstratives," in H. Parret and J. Bouveresse (eds), *Meaning and Understanding* (De Gruyter, 1981) and *The Varieties of Reference* (Clarendon Press, 1982), in which they attributed the notion to Frege.

According to *de-re*-sense theorists there are classes of terms which, when used, express *de re* senses, if they have content at all. If an occurrence of a term in such a class does not denote, then it does not have a sense, and hence is contentless. The most plausible candidates for terms, occurrences of which express *de re* senses, are demonstratives and indexicals. For instance, according to Evans (1982, ch. 6), grasp of the sense of an occurrence of a demonstrative, such as 'this.' is constituted in part by the existence of an "information link" between the denotation of that occurrence of 'this' and the subject. If an occurrence of 'this' does not denote, then there is no such information link, and hence nothing that would count as grasping the sense of that occurrence. But a sense which is such that there is nothing for it to be grasped is not a sense at all. Hence, if an occurrence of "this" does not denote, then according to Evans's account, it lacks a sense, and so is contentless.

Deflationary conception of truth: The view that the predicate 'true' does not stand for any real property, but, excepting its use in indirect endorsements (e.g., 'Fermat's last theorem is true') or compendious ones (e.g., 'Everything he said is true'), is no more than a device of Disquotation (q.v.).

Degree of truth: If two things are borderline cases of red, but one is redder than the other, the redder one is, intuitively, a better candidate for being something of which "red" is true. One might try to capture this idea in terms of "red" being more true of the redder object than of the less red. This gives one an ordering. One might try to use this, along with further data (e.g., comparisons between closeness of pairs in the order) to create a scale. Finally, one might find end-points, to correspond to definite truth and definite falsehood. The points on such a closed scale are degrees of truth. Applied to vagueness, a degrees-of-truth theory does good justice to our feeling that a vague statement is not completely true, but encounters many difficulties, including ones relating to the assignment of degrees to complex sentences, and ones relating to higher-order vagueness.

Denoting versus naming: see **Referentialism**

Describing versus referring: see **Referentialism**

Direct reference: 1. Say that a term (or its use) is *Fregean* if its reference at each possible situation s is what is presented or described, at s, by the term's sense (or by some descriptive condition which the term supplies). A *directly referential* term is one whose uses are *non-Fregean*, since the term's referent, at each situation, is semantically constrained to be the actual referent. David Kaplan, who introduced the terminology, argued that demonstratives and proper names are devices of direct reference. Note that, on this use, a directly referential term may make a "descriptive contribution" to what a sentence says, so long as that contribution is truth-conditionally irrelevant.

2. On another usage, a directly referential term is one whose sole contribution to what a sentence says is its referent; sentences with such terms express singular propositions. Direct-reference accounts of propositional-attitude talk take demonstratives and names to be thus directly referential, and take attitude ascriptions to ascribe (only) relations to the propositions determined by their complements.

Disjunction problem: The disjunction problem is a problem for crude causal accounts of reference and truth-conditions. On a crude causal account the truth-condition of a belief-type B is the type of states of affairs that cause B. The difficulty is that if the state of affairs S causes the belief B, and the state of affairs S* causes the belief B, then the disjunctive state of affairs S or S* causes B. That is, according to the theory, B's truth-conditions are the disjunction of all its causes. This makes error impossible. Naturalistic semantic theories attempt to deal with the problem by specifying a distinction between those causes that constitute B's truth-conditions and those that don't.

Disquotation: According to the disquotational principle for the predicate 'true,' a true biconditional results whatever sentence is substituted for the variable 'P' in the scheme:

'P' is true if and only if P

If it is granted that truth is properly predicated of sentences (as distinct from the thoughts or propositions which sentences may be used to express), and provided that the sentence quoted in the left-hand component of the biconditional is taken to be (a quoted version of) the sentence used in the right-hand component, it appears indisputable that this principle captures a fundamental feature of the notion of truth, since the principle, so understood, appears to be no more than a metalinguistic version of what is usually called the Equivalence Thesis: It is true that P if and only if P. An associated – but much more controversial – thesis is that the predicate 'true' does not stand for a genuine property, and is little more than a device of disquotation, needed only for the purpose of indirect or compendious endorsement (as in 'Pythagoras's Theorem is true' and 'Whatever Aristotle says is true,' respectively) – this is the Deflationary conception of truth (q.v.). Closely related is the famous Redundancy Theory, according to which there is no more to (the concept of) truth than is involved in accepting all instances of the disquotation principle, or the Equivalence Thesis.

Effectively decidable: A statement is effectively decidable if there is a routine procedure which can be followed in any given case and which is guaranteed to lead to a correct decision as to the statement's truth-value. Derivatively, a predicate is said to be (effectively) decidable if there is a routine procedure for determining, with regard to an object (or sequence of objects, in case of an n-place predicate for $n \geq 2$) whether it satisfies that predicate – equivalently, whether the predicate is true of that object (or sequence of objects). Effectively decidable statements are, by their very nature, incapable of being undetectably – or Evidence-transcendently (q.v.) – true.

Epistemic conception of meaning: In one usage, this denotes the view that to understand a word or sentence is to know rules (i.e., conventions) which govern its role in our language practices, in particular of assertion and inference. Rules governing the use of words take the form of conventions for introducing and eliminating terms in a language; they in turn determine when a sentence is assertible and what is inferable from it.

In another distinct, but perhaps related usage, an epistemic conception of meaning is any view which insists that sentence meanings be given in terms of conditions whose satisfaction is always, at least in principle, a recognizable matter – such as conditions of warranted assertion, rather than Evidence-transcendent (q.v.) truth.

Epistemic conception of vagueness: This is the view that vague statements are either true or false (and not both) in borderline cases, although we cannot know which (see **Bivalence**). Similarly, on this view, the major premise of a sorites paradox, for example, 'For all n, if n + 1 grains make a heap then n grains make a heap,' is falsified by some number n, although we cannot know which. The epistemic view permits the retention of classical logic, truth-conditional semantics, and Disquotational principles about truth (q.v.) for vague languages. It is not implied that every unknowable truth is vague. Rather, the view identifies vagueness with a particular kind of unknowability whose origin is conceptual.

Epistemic possibility/necessity: Modal words ('possible,' 'necessary,' 'may,' 'must,' etc.) are sometimes used to express claims about what is possible, or necessary, given our state of knowledge or information. Thus when we say of some acquaintance whose intentions are as yet unclear to us, and are perhaps not yet determinate, that she may come to the party, we are probably not merely claiming that it is a bare logical possibility that she will come, but that we know of nothing from which it can safely be inferred that she will not come. If so, our claim is one of epistemic possibility. A correlative use of 'must' to express epistemic necessity is perhaps exemplified when we say, when the person we are expecting to meet conspicuously fails to alight from the train, that she must have missed it. This use of modal words needs to be carefully distinguished from their use to express other kinds of necessity and possibility, since what is epistemically possible may well not be, for example, logically or metaphysically possible. Thus it is plausible, at least, that whichever of Goldbach's Conjecture (that every even number greater than 2 is the sum of two prime numbers) and its negation is true, is necessarily true. But in our present state of knowledge, both the Conjecture and its negation are epistemically possible.

Error theories: An error theory for a given area of discourse maintains that statements belonging to the discourse are aimed at truth, but are systematically false, owing to the failure of some presupposition – usually ontological in character – of the discourse as a whole. Thus an error theory contrasts with certain other types of anti-realist position, according to which statements of the seemingly problematic kind are to be reinterpreted in some way; for example, by reductive translation into statements of some other unproblematic type, or, quite differently, by construing them as having some other than assertoric function (as on the emotivist theory of ethical utterances, according to which they are aimed not at stating facts but at evincing feelings or attitudes). In the classic example, John Mackie maintained that ordinary moral discourse is error-ridden because there do not exist the distinctively non-natural properties or moral facts required for the truth of its statements. Other examples are the view that mathematical statements are uniformly false because their truth would call for the existence of number, sets, or other kinds of

abstract entity (denied by the error theorist, in this case a nominalist); and eliminativist views about 'folk'-psychological discourse.

Essential property: A property *P* is traditionally said to be an essential property of an object *x* if *x* could not lack *P*, or, better, if *x* could not exist while lacking *P*. Without the qualification about existence, no property would be essential to a contingently existing object if possession of the property by an object required the object to exist. Unfortunately, with the qualification about existence, existence itself becomes an essential property, which is not the intent of the traditional notion.

Essentiality of origin: A thesis due to Kripke and sometimes called 'the necessity of origin.' It claims that the origins of things of certain specified kinds are essential to them. For instance, an individual human being may be said to originate essentially from the zygote he or she in fact originated from. The thesis can also be applied to more abstract entities to which the notion of origin is applicable, such as species.

Euthyphro contrast: The Euthyphro contrast, prominent in a good deal of recent discussion of realism and opposed positions, and especially in the work of Crispin Wright, concerns whether our best judgments in a given area are to be regarded as tracking an independently constituted realm of facts (the realist view) or whether, rather, we should view truth for the discourse's statements as somehow determined by, or constituted out of, our best judgments (the anti-realist option). The label recalls Plato's dialogue, which has Plato maintaining that pious acts are thought to be so by the gods because those acts are pious, while Euthyphro contends for the opposed view, that pious acts are so because the gods take them to be so. Realist and anti-realist may be presumed to agree that there will be a coincidence between the facts of the matter and our judgments made under optimal conditions. The issue then concerns the direction of dependence: are such judgments true because they match up with independently constituted facts, or are those facts themselves no more than a reflection of our best judgments? (See also **Cognitive command**, **Wide cosmological role.**)

Evidence transcendence: To hold that statements of some kind may be evidence-transcendently true or false is to claim that such statements may have determinate truth-values without its being possible, even in principle, for us to discover what those truth-values are. It is standardly taken to be a mark of a certain type of realism with respect to a class of statements to maintain that those statements may be evidence-transcendently true, or false. A realist of this kind holds that the meanings of (at least some of) our sentences are to be given in terms of truth-conditions, the satisfaction of which is a potentially evidence-transcendent matter. The opposed anti-realist contention is that the only notions of truth and falsehood which we may justifiably employ are evidentially or epistemically constrained notions according to which there is an essential connection between truth-values and evidence (and hence between sentence meaning/truth-conditions and evidence). (See also **Bivalence.**)

Expressivism: An expressivist treatment of a given region of discourse maintains that, whatever the surface grammatical form of its characteristic utterances may suggest to the contrary, those utterances are not genuinely assertoric or descriptive, aimed at conveying truths concerning a certain subject-matter, but are instead properly to be understood as serving to express feelings or attitudes. A classic example is the logical-positivists' treatment of ethical utterances – anticipated by David Hume and generally known as the emotive theory – as serving to evince feelings of moral approval or disapproval.

Extension/Intension: *Extension* is a generalization of the notion of reference. Standardly, the extension of a singular term is its referent; of an adjective, noun, or verb, it is the collection of things of which it is true; of a sentence, it is its truth-value. Extensions can be assigned to other meaningful expressions, such as quantifiers.

Some languages are *extensional*. Substitution of expressions with the same extension doesn't change the extension of the whole. (Interpreted) versions of first-order logic are examples: Replacing term *t* with term *s* in a sentence can't change the sentence's extension (truth-value) in such languages, if *t* and *s* have the same extension (referent); likewise, interchange of co-extensive predicates cannot alter a sentence's truth-value. Natural languages certainly don't *seem* extensional (though some have claimed otherwise): 'centaur' and 'unicorn' have the same extension; 'wanted to see a centaur' and 'wanted to see a unicorn' do not.

Intension is used in a variety of ways. A standard use identifies an expression's intension with some aspect of meaning which determines extension. A technical but important use is from possible-worlds semantics, where expressions receive extensions relative to "possible worlds"; an expression's intension is the rule or function assigning its extension at each world. Modal languages are typically intensional – the intension of an expression being determined by the intensions of its parts – but not extensional. David Lewis, Richard Montague, and Robert Stalnaker have identified propositions and properties with possible-worlds intensions.

External realism: see **Metaphysical realism**

Externalism: Externalism is the view that semantic properties of at least some concepts and thoughts, especially the properties of making reference to certain entities and of having certain truth-conditions, do not supervene on intrinsic properties of mental or neural states. Arguments for externalism are mostly appeals to intuitions concerning thought-experiments in which thinkers are supposed to be identical with respect to their intrinsic properties, but are in different environmental contexts: see Hilary Putnam, *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge University Press, 1975), pp. 223–229. If externalism is true then the facts in virtue of which the semantic properties of concepts and thoughts are instantiated must involve relations between the thinker and her environment.

Facts: Our conception of a fact is ambivalent, in a philosophically confusing way. On the one hand, facts belong to the world and are not of our making (except in so far as we have made the world as it is: the fact that there is coffee here is due to something I did). In the traditional version of the correspondence theory of truth, facts in the world are what true propositions correspond with. On the other hand, many people think the fact that Caesar crossed the Rubicon is a different fact from the fact that the conqueror of Gaul crossed the Rubicon, even though there must be a sense in which only one thing happened. There is a case for saying that ordinary language encourages us to individuate facts as finely as we individuate statements or propositions; hence Strawson's ("Truth," *Proceedings of the Aristotelian Society*, suppl. vol. XXIV (1950) pp. 129–156) thought that facts are just what true statements state. This can mislead people into thinking that, since facts are individuated by the concepts we use to express them, the reality to which true statements correspond cannot be fully independent of how we think about it and the concepts we employ. It might be better to give up calling that reality "the facts."

First-order languages: By a first-order language is meant, primarily, a formal language whose logical vocabulary comprises sentential connectives (usually negation, conjunction,

disjunction, the conditional, and the biconditional), together with quantifiers binding just individual- or name-variables. It is the latter condition which determines the order of the language. Higher-order languages contain, additionally, quantifiers binding variables ranging over entities of other kinds, such as properties and relations. Thus '$\forall xFx$' and '$\exists x \exists y Gxy$' might be sentences of a first-order language, since they involve only quantification over individuals; but '$\forall F \forall x(Fx \vee \neg Fx)$' must belong to a language of (at least) the second order, since it involves quantification over properties of individuals. A language will be second-order if it permits quantification over properties and relations of individuals, but not quantification over entities of any 'higher' type, such as properties of properties of individuals, and so on. By a natural extension, 'first- (or second-, etc.) language' may be used to refer to natural languages whose logical vocabulary does not exceed that of first- (or second-, etc.) formal languages.

Frege's Puzzle: Frege's Puzzle is this: How can (uses of) sentences which differ only by terms referring to the same thing differ epistemically (for example, in how informative they are)? This is a puzzle for anyone who believes (a) what a sentence use says is individuated in terms of what the user speaks of in using the sentence (so that what 'Loetze wrote' says turns on the reference, not the sense, of 'Loetze'); and (b) that the epistemic properties of a sentence supervene on what it says. Frege's puzzle is the occasion for Frege's introduction of the notion of sense; it is seen as a major embarrassment for contemporary "direct reference" (q.v.) accounts of assertion.

Full-blooded vs. modest theory of meaning: Contrast introduced by Michael Dummett. A modest theory of meaning for a language is, as he puts it in *The Seas of Language* (Oxford University Press, 1993), p. viii, not intended to "convey the concepts expressible" in it, but to "convey an understanding of that language to one who already had those concepts." A full-blooded theory should also specify what it is for a speaker of the language "to possess the concept it expresses."

Generality Constraint: The Generality Constraint is a version of the principle of compositionality. Applied to linguistic understanding, it entails that the ability to understand, say, a simple subject-predicate sentence Fa is composed of distinct abilities, the ability to understand a and the ability to understand F. These abilities must, furthermore, be manifested in the understanding of other sentences involving a and F.

A more exact characterization of the Generality Constraint (as applied to linguistic understanding) is as follows. Let e be an expression. If someone understands e, then that person possesses the ability to understand every sentence S which results from placing e into the open position of some linguistic string L which satisfies the following conditions (1) that person understands all the expressions in L (and the person has a mastery of the relevant ways of composing the elements of the resulting sentence); (2) the result is syntactically and semantically well-formed; and (3) the result is not too complex to be processed by the speaker.

For example, suppose it is claimed that someone understands the predicate, 'feels pain.' If the Generality Constraint is true, then that person must also have the ability to understand sentences such as "I feel pain" and "Bill feels pain," if she grasps the first-person pronoun, the name "Bill," and predication. (Thus, the possibility that someone has an understanding of their own mental self-ascriptions, and third-personal physical ascriptions, but no understanding of third-personal mental ascriptions, is inconsistent with the Generality Constraint.)

Traditionally, the Generality Constraint is formulated as a prerequisite for possessing the ability to entertain thoughts, rather than in terms of linguistic understanding.

The Generality Constraint was first introduced and put to use in chapter 3 of Strawson's *Individuals* (Methuen, 1959), though it is implicit in many defenses of the compositionality of language and thought.

Higher-order vagueness: The word 'heap' has first-order vagueness because it can have borderline cases; it has second-order vagueness because the expression 'borderline case of "heap"' can itself have borderline cases. More formally, suppose that there is a standard way of constructing a metalanguage for any given language, in which any vagueness in the latter can appropriately be described. Given a language L, inductively define a sequence of languages L_1, L_2, L_3, … by letting L_1 be L itself and L_{n+1} be the metalanguage for L_n. Then L is nth-order vague if and only if L_n is vague in the ordinary sense; L is higher-order vague if and only if it is nth-order vague for some $n \geq 2$. Corresponding notions of vagueness can be defined for individual expressions. It is plausible that all natural languages are nth-order vague for all n. Theories of vagueness often have difficulty in accommodating higher-order vagueness because they treat the metalanguage for a vague language as though it were precise.

Homophonic specification of meaning or of reference, or of truth-conditions: When the object-language (containing the expressions whose meaning/reference/truth-conditions are to be given) forms part of the metalanguage (in which the specification is to be effected), those meanings, and so on, may be specified homophonically, by using those very expressions themselves. Thus:

'cat' means cat
'mice' refers to mice
'cats eat mice' is true if and only if cats eat mice.

Hyperintensional: 1. Properties and propositions (thought of as what predicates and sentences express) are sometimes identified with constructions out of things such as possible worlds. *Hyperintensional* theories hold that propositions (and/or properties) are basic entities, not reducible to constructions from worlds or the like, and that they are more fine-grained than constructions from possible worlds (since, for instance, predicates can express different properties although they have the same intension).

2. Linguistic contexts in which expressions with the same (possible-worlds) intension are not inter-substitutable *salva veritate* are sometimes called *hyperintensional*. Propositional-attitude contexts are the paradigm of such linguistic contexts.

Implicature: Speakers often communicate something other than what they strictly and literally say. For example, a speaker might use the words, "There is a restaurant nearby" in order to say that there is a restaurant nearby and might thereby communicate, in addition, that they expect the restaurant to be open. Grice drew attention to the importance of this distinction between what is strictly and literally said (more generally, basic speech acts) and other things communicated by saying it (more generally, non-basic, or derivative, speech acts), and used the label *implicatures* for what is communicated without being strictly and literally said. In addition to drawing attention to these distinctions, Grice aimed to account for the ways that implicatures can be generated and discerned. As part of that project, he also distinguished between what he called *conversational implicatures* – implicatures that can be discerned on the fly by exploiting general principles governing rational communication – and *conventional implicatures* – which may have originated as conversational implicatures, but now need to be learned by rote.

Indeterminacy of translation: The thesis, associated with Quine, that the methods of Radical translation (q.v.) do not uniquely determine a single translation from one language to another. In other words, no matter how much linguistic–behavioral data radical translators of a given language have to go on – indeed, even if they had access to all such data available in principle – they could still arrive at distinct manuals of translation which, while making optimum sense of the native speakers' linguistic behavior, were in conflict with each other about the correct translation of particular expressions. The thesis of the indeterminacy of translation proper maintains that this conflict can take the form of a divergence over the *truth-conditions* of certain sentences, so that one optimal manual of translation may represent a particular utterance as saying something true, while another represents it as saying something false. In a weaker form, however, the thesis is that the assignment of reference to the constituents of a sentence – singular terms and predicates, for instance, occurring within it – can vary, even though the truth-conditions of the sentence be fixed (see **Inscrutability of reference**).

Indexicals: Indexicals are words such as "I," "you," "here," and "now." The designation of such words (what they "stand for") shifts from use to use, depending on various contextual factors. The designation of "I" depends on who the speaker is; the designation of "you" on the intended audience; the designation of "now" on the time, and so forth. Demonstratives such as "this" and "that," and demonstrative phrases such as "that man" or "this computer," are usually reckoned to be a subclass of indexicals.

Indiscriminability: To discriminate between *x* and *y* is to recognize a difference between *x* and y. Thus *x* and y are indiscriminable if and only if no difference is recognizable between them. Things may be indiscriminable in one respect and not in another (e.g., in color but not in shape), by one means and not by another (e.g., by chemical analysis but not by touch), by one person and not by another, when presented in one way and not when presented in another, and so on. In general, indiscriminability is a reflexive and symmetric, but not transitive relation: each term in a series may be indiscriminable from its neighbors, even though the first and last terms are easily discriminable. Sorites paradoxes (see Chapter 28, SORITES) result from the assumption that whenever two things are indiscriminable, an observational term correctly applicable to one must also be correctly applicable to the other.

Individual essence: Intuitively, the individual essence of a thing is those features of it which in some sense make it what it is, or constitute its identity. A collection of essential properties is an individual essence of an object *x* if it is not possible for any object other than *x* to possess them all. However, this definition is consistent with a thing's having more than one individual essence, and so does not entirely capture the traditional notion.

Informational theories of truth-conditional content: According to informational theories, the truth-condition of a belief-state B is the information that B carries, or would carry under certain circumstances: see F. Dretske, *Knowledge and the Flow of Information* (MIT Press, 1981). Different accounts result from different specifications of the circumstances in which information determines truth-conditions. For example, optimal-conditions theories specify the conditions as ones which are epistemically optimal: see R. Stalnaker, *Inquiry* (MIT Press, 1984), J. Fodor, "Psychosemantics, or where do truth conditions come from?" in W. Lycan (ed.), *Mind and Cognition*, (Blackwell, 1990). These are conditions in which B is tokened if and only if it is true. Certain teleological theories specify the truth-condition-constituting information as the information that it is B's biological function to carry: see R. Millikan, "Biosemantics," *Journal of Philosophy*, 86 (1989) and D. Papineau, *Philosophical Naturalism* (Blackwell, 1993). The accounts so far

devised seem either to appeal to intentional specifications of truth-condition-constituting circumstances or to fail to assign determinate truth-conditions.

Inscrutability of reference (see also **Indeterminacy of translation**)**:** The thesis that the truth-condition of a sentence, and hence its truth-value across all possible worlds, may be held constant consistently with variations in the assignment of reference to its constituents, in particular to the singular terms and common nouns which it may contain. That the thesis holds is the gist of Quine's famous 'gavagai' argument, and is ostensibly established in a very general form by Putnam's permutation argument (see **Permutation argument**).

Intension: see **Extension/Intension**

Intensionality: A context is intensional just if its truth-value is liable to be altered by the substitution within it of expressions which have the same Extension (q.v.) as those for which they are substituted. The principal examples of such contexts are modal sentences, sentences ascribing intentional attitudes (see **Intentionality**), and contexts of direct quotation. Thus, it is necessary that nine is greater than seven, but not necessary that the number of planets is greater than seven, although "nine" and "the number of planets" have the same extension (i.e., reference). Again, someone may believe that all lions have hearts without believing that all lions have kidneys, although "has a heart" and "has kidneys" have the same extension (i.e., are true of the same things). Likewise, 'Richard was called "Cœur de Lion" because of his bravery' may be true, while 'Richard was called "Richard I" because of his bravery' is doubtless false, although "Cœur de Lion" and "Richard I" have the same reference.

Intention-based semantics: Intention-based semantics is a term that has been used by Stephen Schiffer to refer to Grice's analysis of meaning. The title captures the centrality given to the concept of intention in this analysis. This title also brings out the way in which Schiffer once used Grice's work as part of a larger program of reducing semantics to psychology and psychology to the physical-cum-functional.

Intentionality: That characteristic of mental states which consists in their being *about* something, in their being directed upon some particular (putative) object or state of affairs. The term was coined in this technical sense by the Scholastics (from the Latin *intendo*, "to point") and revived by Brentano, who held that intentionality defines the distinction between the mental and the physical, and constitutes a decisive barrier to any kind of reduction of the former to the latter. Brentano held, more specifically, that the intentionality of the mental shows itself in two characteristics: mental states may be indifferent to the non-existence of their objects – the mere fact that the Holy Grail does not exist is no barrier to Gawain's being correctly described as hoping to find it – and they are typically sensitive to variation in the mode of presentation of the object they concern – Lois Lane may hope to marry Superman, but not hope to marry Clark Kent, for instance, even though they are one and the same.

Derivatively, certain kinds of psychological sentences are described as intentional when they exhibit corresponding linguistic features, that is, do not sustain widescope existential generalization and are prone to the kind of substitution failures associated with Intensionality (q.v.). Ascriptions of propositional attitude – belief, desire, hope, fear, intention, and so on – are paradigms of such intentional contexts.

Interpretation: In general terms, an interpretation of a language is an assignment of meanings to the expressions of the language. In the context of the semantic study of formal languages employed in logic, the term 'interpretation' bears a somewhat more precise

meaning, which may be illustrated here for the case of a First-order language (q.v.) with logical vocabulary comprising the usual sentential connectives and quantifiers binding individual variables. Giving an interpretation, then, consists in specifying a (non-empty) set – the domain of the interpretation – and, on this basis, assigning references to the various items of primitive non-logical vocabulary, which will always include a selection of predicates, and may also include functional expressions and individual constants. The general idea is that these assignments should be made in such a way as to induce, via the fixed meanings of the logical expressions, truth-values on the sentences of the language, which are then said to be true, or false, relative to that interpretation. The individual variables range over the specified domain, that is, they may take any of its elements as values. Each 1-place predicate is assigned a subset of the domain – intuitively, the objects in the domain of which it is stipulated to be true on that interpretation; 2-place predicates are assigned sets of ordered pairs of elements of the domain; 3-place predicates sets of ordered triples, and so on. Individual constants are assigned elements of the domain, and functional expressions are assigned functions taking objects in the domain as arguments and yielding objects in the domain as values. Thus a simple sentence 'Rab' will be true under an interpretation I if and only if the ordered pair of elements of I's domain assigned to the constants 'a' and 'b' respectively belongs to the set (of ordered pairs) assigned to the 2-place predicate 'R.' The universally quantified sentence '∀xFx' will be true under I if and only if 'Fx' is true no matter which element of I's domain is taken to be the value of the variable 'x' (i.e., iff every element of the domain belongs to the set assigned to 'F'). A conjunction of sentences 'A & B' will be true under I iff both 'A' and 'B' are separately true under I. And so on.

In terms of these basic ideas, certain further important semantic concepts may be defined. For example, an interpretation I is said to be a model of a set of sentences Γ iff every sentence in Γ is true under I. An inference Γ ⊢ A is valid iff every model of Γ is also a model of {A}. A sentence A is logically true iff every interpretation of A's language is a model of {A}.

Kinds: Kinds are Universals (q.v.), having Particulars (q.v.) as their instances. Terms denoting kinds belong to the more general categories of Sortal terms (q.v.) and Mass terms (q.v.). Examples belonging to the former category are 'tiger' and 'lemon,' while examples belonging to the latter are 'gold' and 'water.' All of these are examples of *natural kind* terms, which are to be contrasted with terms for *artifactual* kinds, such as 'pencil' and 'yacht.' A distinguishing feature of natural, as opposed to artifactual, kinds is that they are typically subjects of natural scientific law (for example, it is a natural law that gold consists of atoms containing 79 protons in their nuclei). It is nowadays widely held that natural kind terms are Rigid designators and consequently not definable by means of complex descriptions in the way that empiricist philosophers such as John Locke believed them to be.

Law of Excluded Middle: The law claims that every instance of 'P or not-P' is true, or is a theorem. This is guaranteed by Bivalence (q.v.), given that 'not' toggles truth-values. However, it can hold without bivalence, for example in a three-valued system in which 'not' turns any non-truth into a truth. Among theories of vagueness, supervaluation theories maintain the Law of Excluded Middle, while departing from bivalence: some sentences are neither true (that is, true-on-all-valuations) nor false (that is, false-on-all-valuations); but every instance of 'P or not-P' is true-on-all-valuations and so true.

Leibniz's Law: Leibniz's Law is the principle that if *a* is identical with *b* every property of *a* is a property of *b*. It is what distinguishes *absolute* from any form of *relative* identity.

There are many apparent counter-examples involving psychological and modal properties, but these cannot be genuine counter-examples, since Leibniz's Law is definitive of absolute identity. Thus Leibniz's Law must be sharply distinguished from the false principle of substitutivity that if '*a*' and '*b*' are two singular terms for the same object, then replacement of '*a*' by '*b*' will be possible *salva veritate* in any context. Leibniz's Law is not in formal conflict with the Relative Identity Thesis (q.v.) of Peter Geach, since the latter entails not that Leibniz's Law is false, but that it is inexpressible, since no language can contain an expression for absolute identity (q.v.).

LF: LF is a hypothesized level of syntactic structure that encodes features traditionally captured by logical systems, such as variable binding and scope relations between operators. Thus, initially, LF was an acronym of logical form. LF, however, is a syntactic structure and so hypotheses about LF are grounded in otherwise accepted syntactic relations and operations. For example, the scopal properties of quantifiers and other operators are encoded at LF in terms of the movement or displacement of the relevant phrases to higher positions such that the moved phrase scopes over the minimal structure that includes it, with the launch site of movement treated as a bound variable. Crucially, such movement operations are otherwise witnessed throughout language, so the hope is that the particular properties at issue, such as scope, fall under already understood properties of syntax. More generally, LF is hypothesized to be the interface between syntactic structure and semantic interpretation. It remains open, however, just what properties of semantic interpretation are fixed by LF. Thus, LF is not the same notion as logical form (q.v.). LF as classically conceived is a controversial hypothesis, both empirically (not all scope-taking can be understood as movement) and theoretically, due to developments in so-called level-free approaches to syntax.

Locutionary/Illocutionary/Perlocutionary acts: This threefold distinction was adopted by Austin after he abandoned his constative–performative distinction (see **Performative utterances**). A locutionary act is the act of saying something, characterized by Austin as a matter of uttering certain words "with a certain 'meaning' in the favourite philosophical sense of that word, i.e. with a certain sense and with a certain reference": see Austin, *How to Do Things with Words* (Clarendon Press, 1962), p. 94. An illocutionary act is the performance of an act *in* saying something in this sense, such as giving a warning, making a promise, and so on. A perlocutionary act is an act – such as drawing someone's attention to a bull or reassuring them – performed *through* or *by* the illocutionary act.

Logical form: The logical form of a sentence is a formula framed in a logical system (e.g., first-order logic) associated with the sentence. The purpose of the association is to explain the truth-conditions of the sentence (in abstraction from word meaning) and what relations of entailment the sentence stands in. The precise nature of the association is controversial, but a shared conception is that the form may and typically does depart from the apparent grammatical organization of the sentences at issue. Thus, for example, Russell's theory of definite descriptions renders the logical form of sentences of the kind *The F is G* as consisting of three conjuncts, and free variables bound by existential and universal quantifiers. The proposal explains, among other things, why sentences of the target kind may be ambiguous under negation in the way sentences of the kind *N is G* are not (where *N* is a proper name). Similarly, Davidson's (initial) account of 'action sentences' involving transitive verbs renders their logical form in terms of a ternary relation involving a bound variable ranging over events. The analysis explains, among other things, the semantic properties of sentences such as *Bill hit Garry and he didn't like it*,

where the pronoun *it* picks up on the event of the hitting. The general idea of a logical form is widely accepted, even though many disputes remain as to whether the association between logical form and the target sentences is cognitively grounded, a descriptive or theoretical generalization, or a mere paraphrase.

Löwenheim–Skolem Theorems: These are some theorems in the model theory of first-order logic, established originally in several versions by Leopold Löwenheim and Thoralf Skolem, which disclose limitations on the expressive capacity of first-order languages. The Downward Löwenhiem–Skolem Theorem asserts that if a set of first-order sentences has a model at all, that is, an Interpretation (q.v.) in which all the sentences in the set come out true, then it has a countably infinite model (i.e., a model whose domain has exactly as many elements as there are natural numbers). The Upward Theorem asserts that if such a set of sentences has a model in any infinite cardinality, it has models in every infinite cardinality. The theorems have played a central role in one of Hilary Putnam's arguments directed against the position he calls external or Metaphysical realism (q.v.).

***Malin génie*:** The point of the *Malin génie* in Descartes's first *Meditation* is to introduce the most extreme skepticism possible. Some recent philosophers replace the *génie* with a scientist who keeps my brain in a vat, controlling it and all its inputs. The point is the same. No argument could show that I am not being deceived by a *génie*, because the *génie* could make me think arguments sound when they are not; so could the scientist. One can try to undermine the skeptical hypothesis, but one will get nowhere by arguing (as some have) that at least the *génie*/scientist must know some truths; for the *génie*/scientist is only a dramatization of the possibility that our belief-system wholly fails to match the world. The coherence theory of truth attempts to get round it by suggesting that our belief-system, or an idealization of it, determines how the world is. If that were so, there could be no general failure of match. Of course the doubt could still be raised that any particular beliefs (e.g., my present ones) might be mistaken – even my belief that they cohere may yet be false. Perhaps that doubt should not worry us.

Manifestation argument: Common term for an objection (associated particularly with Michael Dummett) to any semantic theory which allows the nature of speakers' private states (see **Privacy**) to influence the meaning of the words they use: it cannot explain how a language-learner could ever show their competence in the language to be learnt. For competence would involve associating the right kind of private state with the words of the language, and whether they are doing this or not can never (by the definition of a private state) be known to anyone else. The manifestation argument is a close relation of the arguments from Acquisition (q.v.) and Communicability (q.v.). (See Chapter 11, MEANING AND PRIVACY.) Interest in these arguments has been aroused by Dummett's suggestion that they support an 'anti-realist' approach to semantics, shifting emphasis away from the conditions under which a sentence is true, perhaps Evidence-transcendently (q.v.) towards those under which a speaker might properly assert it.

Mass terms: Mass terms, such as 'water' and 'gold,' differ from Sortal terms (q.v.) in being dissective: gold is divisible into parts which are themselves gold, unlike the parts of a horse, which are not themselves horses. They also differ from sortal terms in not supplying principles of enumeration for their instances. Thus, whereas it makes sense to ask *how many* horses there are in a field, it only makes sense to ask *how much* water or gold there is in a room. Even so, mass terms do clearly have Criteria of identity (q.v.) governing their use, for it makes sense to ask whether the gold now in this room is the *same* gold as was in this room yesterday.

Meaning-truth platitude: This is the (allegedly platitudinous, and certainly widely endorsed) principle that 'the truth-value of a statement depends only upon its meaning and the state of the world in relevant respects,' a formulation which, along with the term, comes from Crispin Wright's "Kripke's account of the argument against private language," *Journal of Philosophy*, 81:12 (1984), pp. 759–778. The principle plays a central role in his argument that Semantic irrealism (q.v.) leads to an untenable global irrealism. (For discussion, see Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, §3.)

Mentalese: Mentalese is the name of the hypothesized language in which we think. According to advocates of the existence of Mentalese, for example, J. Fodor, *The Language of Thought* (Harvester, 1976), to think a thought, form a belief, or remember a fact, is to produce, in one way or another (which way depending on whether it is a belief, a memory, etc.), a sentence in an internal language. The existence of Mentalese explains the semantic properties of mental states in terms of the semantic properties of their component Mentalese expressions. This language, unlike natural languages, does not possess its semantics in virtue of the intentions or conventions of its users. Semantic naturalists think that the expressions of Mentalese possess their semantic properties in virtue of natural relations, especially causal and nomological relations, to extra-mental items. The Mentalese hypothesis is usually defended as being part of the best explanation for certain features of thought. The two most prominent features are productivity (a thinker can think complex thoughts) and logical inference: J. Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press, 1987).

Metaphysical realism: This term, along with 'external realism,' is employed by Hilary Putnam to refer to an amalgam of several closely associated philosophical ideas about the relations between language and reality, and between truth and knowledge or justifiable belief. One component on which Putnam places considerable emphasis is that even an ideal theory (a theory that is '*epistemically* ideal *for humans*,' ideal by the lights of the operational criteria by which we assess the merit of theories) may nevertheless be, in reality, false. More commonly, Putnam characterizes metaphysical realism in terms of three other theses, of which he takes this feature to be a consequence: that 'the world consists of a fixed totality of mind-independent objects,' that 'there is exactly one true description of the way the world is,' and that 'truth involves some sort of correspondence between words or thought-signs and external things and sets of things.' Putnam attacks this kind of realism, advocating instead what he terms 'internal realism.'

Millian semantics (Direct-reference theory): A Millian, or direct-reference, semantic theory is a hypothesis according to which the semantic value of a proper name (the contribution that it makes to the determination of the content of utterances containing it) is simply the referent of the name. According to this hypothesis, the thought or proposition expressed by a sentence containing a proper name is determined as a function of the individual named, perhaps by containing that individual as a constituent. The contrasting hypothesis is that reference or denotation is mediated by a sense or connotation. According to the contrasting theory, the semantics for the name provides some kind of abstract object – a sense or mode of presentation – which is a constituent or determinant of the thought or proposition expressed, and which also determines a referent.

Minimal truth: The term 'minimalism' is sometimes used to refer to a doctrine about truth not readily distinguishable, if indeed distinct at all, from the Deflationary view (q.v.). However, the term has also been employed (especially in recent work by Crispin Wright, and discussions that this has generated) for a characterization of truth intended to be

neutral as between realists and their various opponents. According to this account, it suffices for a predicate T to be or function as a truth-predicate that it satisfy the Disquotation Scheme: '"P" is T if and only if P' and that it exhibit certain features embodied in or derivable from certain 'platitudes,' as Wright describes them: centrally, that to assert a statement is to present it as true, and that any truth-apt content has a significant negation which is likewise truth-apt.

Modal realism: There are two quite distinct doctrines about modality to which the label 'modal realism' may be applied. Modal realism, in one reasonable sense of the term, is the view that there are irreducible modal truths or modal facts expressible by sentences featuring modal operators or equivalent devices, such as 'Necessarily 17 is prime,' 'Uncles cannot but be brothers,' and so on. Realism in this sense is opposed by any view which denies the existence of a distinctive class of modal facts, either by rejecting modal talk altogether, or by giving a reductive account of it, or by accepting such talk as irreducible but arguing that, properly understood, it has some quite different, non-fact-stating role, as on a Non-cognitivist view (q.v.). In a quite different sense, 'modal realism' denotes a view about the existence of possible worlds. In its extreme form, as famously advocated by David Lewis, this second kind of modal realist holds that there literally are many other possible worlds besides the world we inhabit, each such world being spatially and temporally (and therefore causally) isolated from all other possible worlds. More moderate forms of realism about possible worlds have also been defended, most notably by Robert Stalnaker. It is clear that modal realism in the first sense does not require modal realism in the second. Indeed, since possible-worlds realists favor a reductive analysis of modal idioms to quantification over a domain of possible worlds, the two forms of realism appear incompatible. (For discussion, see Chapter 31, MODALITY.)

Model-theoretic arguments: Model theory is the branch of mathematical logic which studies the interpretation of formal languages (see **Interpretation**). Thus a model-theoretic argument might be any argument that draws on the results or techniques of model theory. In the philosophy of language, model-theoretic arguments which appeal to the Löwenheim–Skolem Theorems (q.v.) and to Permutation results (q.v.) have been deployed by Hilary Putnam, W. V. O. Quine, Donald Davidson, and others in support of conclusions about indeterminacy of reference or meaning. (See Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM, for discussion of Putnam's use of these arguments.)

Myth of the museum: A term used by Quine to stigmatize the idea that there can be facts about the meanings a speaker assigns to particular expressions transcending anything that might in principle be determined as a correct translation of that speaker's utterances by a radical translator. (See **Radical translation**.)

Natural and non-natural meaning: The distinction, drawn in these terms at least, derives from Grice, 'Meaning,' *Philosophical Review*, 66 (1957). Natural meaning (meaning_{n}) is the sort of meaning possessed by things in nature such as clouds (mean rain), or smoke (means fire), or wounds (mean damage). Non-natural meaning (meaning_{nn}) is possessed by our words and sentences, and by some of our actions and gestures. Grice draws attention to a number of differences between the two kinds of meaning; these include the fact that while "x means_{n} that p" entails that p, the entailment fails for meaning_{nn}.

Natural kind externalism: A version of Semantic externalism, according to which the contents of natural kind terms such as 'water,' 'gold,' and 'tiger' are partly determined by features of the external world. Such features typically involve underlying properties

shared by members of the relevant natural kind, such as molecular structure or atomic number. The most famous argument for natural kind externalism is Hilary Putnam's *Twin Earth* thought experiment. (See **Semantic externalism**.)

Necessity of identity: A thesis defended by Marcus, Wiggins, Kripke, and other philosophers, according to which, if objects *x* and *y* are the same, they are necessarily the same. Instances may be formed in ordinary language by replacing the variables '*x*' and '*y*' with proper names or demonstratives, as in 'if Superman is Clark Kent then, necessarily, Superman is Clark Kent.' Since Superman *is* Clark Kent (treating the fiction as fact) it follows that it is necessary that Superman is Clark Kent, a surprising result if one expects everything that is necessary to be discoverable *a priori*. Marcus, Wiggins, Kripke, and others conclude that some necessities are *a posteriori*.

Necessity of origin: see **Essentiality of origin**

Non-cognitivism: 'Non-cognitivism' is probably best understood as a generic term for any of a variety of views concerning some specified region of discourse according to which, to the extent that the surface syntax of its characteristic utterances suggest that they are aimed at recording discoveries (or cognitive achievements), this appearance is misleading. So understood, the non-cognitivist's thesis is a purely negative one – to the effect that the characteristic utterances of the discourse in question do not serve to depict or represent possible objects of knowledge – and is therefore compatible with a variety of positive theses about the function of those utterances. In particular, while it is consistent with, it does not require, the adoption of an expressivist or projectivist theory of the discourse (see **Expressivism** and **Projectivism**).

Normativity of meaning: It is a central ingredient in understanding an expression to grasp that there are associated with it conditions for its correct application. Put another way, it is essential to any expression's possessing whatever meaning it does, that there are rules for its correct use. In this sense, meaning is normative. The normativity of meaning is frequently emphasized in the later writing of Wittgenstein, and especially in his discussions of rule-following: see *Philosophical Investigations* (1953), §§143–242; *Remarks on the Foundations of Mathematics* (2nd edn, 1956, part VI). It is a common theme in subsequent discussion, and especially in Saul Kripke's interpretation of Wittgenstein on rules, where certain attempts to explicate the notion of meaning in naturalistic terms (such as dispositional theories) are criticized as being incapable of accounting for the normative aspect of meaning. (See Chapter 25, THE NORMATIVITY OF MEANING).

Ontological conception of vagueness: This is the view that not all vagueness arises in thought and talk; at least some of it arises in what is thought and talked about, independently of its being thought or talked about. On this view, it may be vague what the spatio-temporal boundaries of an object are, whether one object is part of another, whether objects are identical, or whether a given object has a given property. Such borderline cases are held to result in the failure of the corresponding statements to be either true or false. Some technical treatments of vagueness consistent with the Semantic conception of vagueness (q.v.) are also consistent with the ontological conception.

Ontology: Ontology is the branch of metaphysics which is concerned with the study of being in general. As such, it concerns such matters as the nature of existence and the categorial structure of reality. That all entities occupy distinctive places within a categorial hierarchy is an idea traceable at least as far back as Aristotle. Different systems of ontology propose different categorial schemes. For instance, some schemes regard the distinction between concrete and Abstract objects (q.v.) as being most fundamental,

while others accord this status to the distinction between Universals (q.v.) and Particulars (q.v.). Again, some schemes take the category of substance to be more basic than the category of events, whilst others take the converse view. Yet other schemes take 'particularized qualities' or 'tropes' to be the most fundamental category of physical entities. Ontological categories need to be clearly distinguished from natural kinds (see **Kinds**).

Opacity: An opaque construction turns a position open to the substitutivity of identity into one not open thereto. Take the *doubts that* construction. It maps the word 'Alice' and the sentence

(1) Tully was an orator

to the sentence

(2) Alice doubts that Tully was an orator.

The construction seems opaque. For the position of 'Tully' in (1) is open to substitutivity, since from for example

(3) a = Tully

and (1) the sentence 'a was an orator' follows. But 'Tully' in (2) is not so open, since 'Alice doubts that a is an orator' does not follow from (2) and (3). Constructions not opaque are called *transparent*. (Definition and terminology are due to Quine.)

Not all non-extensional contexts are opaque. For example, given that descriptions are quantifiers, and that proper names and demonstratives are rigid designators (and so, if co-referential, can be substituted in modal contexts *salva veritate*), modal constructions such as *it is necessary that S* are not opaque. But they are non-extensional.

As opacity is a property of constructions, its presence depends upon the presence of grammatical complexity. Quotation, for example, does *not* involve opacity if quotation names are spelling names. For then "Bob" is short for

'B' plus 'o' plus 'b'

which does not contain the word "Bob."

Openness and mutual knowledge: Mutual knowledge is knowledge possessed by two or more individuals. It is a form of iterated knowledge where it is held that A knows that p & A knows that B knows that p & A knows that B knows that A knows that p & …. The iteration here is potentially infinite. Despite the iteration involved in the account of mutual knowledge, some philosophers have held that the regress involved is perfectly harmless, and that it is obviously present in many cases of knowledge. Furthermore, it is taken by some to be a condition of knowledge. Some Griceans have employed a condition of mutual knowledge as a way of ensuring the sufficiency of their analysis of meaning. Without some such condition, the analysis is open to counter-examples based on deception. The concept of openness is designed to do the same work as that of mutual knowledge. Without appeal to an infinite regress of knowledge, the openness condition is meant to ensure that the speaker intend that all her intentions are recognized. In this way the deception which threatens the sufficiency of the analysis is blocked.

Parataxis: On Davidson's paratactic account of

(1) Gettier said that Sleigh slept

grammatical form masks semantical form. To utter (1) is to utter the complete sentence 'Gettier said that,' whose 'that' is a demonstrative; its reference is the ensuing utterance of 'Sleigh slept.' The point of the latter utterance is not to assert that Sleigh slept, but merely to provide a referent for 'that'; the force of the whole thing is something like

(2) An utterance of Gettier said-the-same-as *this* Sleigh slept.

Among the virtues of this view are that it absolves sentences such as (1) from the charge that they violate the principle of substitutivity; and it apparently allows an account of such sentences to get by with only extensional semantic values. Among its apparent vices are that, in many cases (e.g., 'each author said that he would autograph his book'), it is implausible that we relate individuals to utterances; it seems, instead, that we relate them to *interpretations* of (possible) utterances.

Particulars: Particulars are normally contrasted with Universals (q.v.), the former being instances of the latter: for example, a particular apple is an instance of the universal *apple*, which in this case is a natural kind (see **Kinds**). Particulars may be concrete objects, as in this case, or they may be Abstract objects (q.v.), such as numbers and sets. However, there may also be particulars which do not seem to qualify as 'objects' or 'Things' (q.v.) at all, in any very robust sense, such as the particular smile on someone's face, or the particular color of this apple. Items like these are sometimes called 'particularized properties' or 'tropes.'

Performative utterances: J. L. Austin's term, designed to capture the fact that we do things with words. Austin distinguished performative utterances from what he labeled *constatives*. Austin once believed that only constatives could be said to be true or false; performatives are, by contrast, felicitous or infelicitous. The utterance of a performative was thought by Austin to be not so much a saying of something as a doing of something. Thus, in his view, to utter the words 'I do' or 'I promise to be there,' in appropriate circumstances, is to bind oneself in marriage or to promise to be there, rather than to state that one is doing so. The precise drawing of this distinction gave Austin much trouble, and he eventually abandoned it.

Permutation argument: A permutation of a set of objects is any one-to-one function mapping that set on to itself. Under certain assumptions about the structure of a language, it can be shown that given one Interpretation (q.v.) or scheme of reference stipulations for the language, and given any permutation of its domain, there is an alternative, 'unintended' interpretation of the language based on that permutation, which makes quite different referential assignments to its singular terms and predicates, but which induces the same truth-values on the sentences of the language as does the given interpretation. This has been taken by some philosophers (especially W. V. O. Quine and Donald Davidson) to imply indeterminacy, or inscrutability, of reference, or (by Hilary Putnam) to raise insuperable difficulties for 'external' or 'Metaphysical' realism (q.v.).

Physical externalism: see **Natural kind externalism**

Physicalism: The thesis that reality is nothing more than physical reality, and hence that physical science can, in principle, give a complete description of all that there is. The doctrine is widespread, and influential in the philosophy of language in so far as widely conceived as enjoining some form of reductive or reconstructive account of those areas

of discourse – about values, or semantics, or intentional psychology, or modality, for instance – whose subject-matter does not appear to be physical in any straightforward sense. Physicalism has been a driving force in Quine's philosophy of language in particular, and a target of Putnam's writing since the mid-1970s.

Platonism: In the theory of meaning, this term is often applied to the view that the meanings of terms and sentences are (1) concepts and propositions which (2) are non-spatio-temporal entities known by non-perceptual intuition.

Possible world: A possible world is a way the world might be, a complete possible situation or state of affairs. The concept of a possible world has diverse formal and philosophical uses, and has been given very different philosophical explanations by different philosophers. In abstract formal semantics, possible worlds are primitive elements that are constituents of the models used to define the semantic values for expressions of the language of modal logic. Philosophers have appealed to them to formulate metaphysical theses and to give philosophical analyses of epistemological, semantic, and metaphysical concepts such as supervenience, counterfactual conditionals, dependence and independence of various kinds, potentiality, essence, information, obligation, and ability. According to some, possible worlds are literally worlds, universes parallel to our own; according to others, they are maximal properties that the world might have, or maximal propositions, or complex abstract structures whose elements are actual individuals, properties, and relations.

Principle of Charity: The Principle of Charity is a supposed important constraint, invoked by Davidson, on the interpretation of the thoughts of others. The Principle says that interpretation must proceed in such a way that the judgments attributed to the others come out, for the most part, as true. It is argued to follow from the idea that we can disagree with someone, that is, identify a thought of hers and label it false, only against a background of substantial agreement. This in turn is said to follow from Semantic holism (q.v.). Some philosophers take the Principle of Charity to be implausibly strong, and do not see why a person or persons should not, for example through the bad luck of being faced with many cases of misleading evidence, end up with beliefs which are largely false. Some philosophers, for this reason, prefer to see interpretation constrained by the Principle of Humanity (q.v.).

Principle of Humanity: The Principle of Humanity is preferred by some (e.g., Grandy and Lewis) to the Principle of Charity (q.v.) as a guiding constraint on interpretation. The Principle of Humanity says that we should interpret others as thinking and saying what we would have thought or said had we been in their circumstances, for example if we had had their sensory equipment, undergone their upbringing, been through their life experiences, and so on.

Privacy: As it is understood in the philosophy of mind, the term 'privacy' is used to mean the property of being knowable to one person only. Certain mental states or events, it can plausibly be thought, can be known only to the person in whose mind they take place; nobody else can know what they are like. Whether there are any private states in this sense, and if so which, and what makes them private, are central questions in the philosophy of mind and the theory of knowledge. In the philosophy of language the main questions about privacy are (1) whether the nature of speakers' private states can affect the meanings of their words, or whether meaning is constituted solely by publicly knowable features of speakers and their behavior; and (2) whether there could be a Private language (q.v.).

Private language: A private language is a language which a person uses to record thoughts about their own private states (see **Privacy**); it is often also understood to be a language which, in principle, only that person can understand. The most famous discussion of private language occurs in Wittgenstein' s *Philosophical Investigations*, §§242–258, the burden of which is that no such language is possible, since it would not satisfy the conditions under which terms can have a meaning, even a private one. Wittgenstein's introduction of the notion of a private language (*Philosophical Investigations*, §243) appears to use both the above definitions simultaneously; this may indicate a (contestable) assumption that the first entails the second.

Projectivism: A projectivist theory of a given region of discourse maintains that in casting its characteristic claims in assertoric or propositional style, we are not properly understood as attempting correctly to describe mind-independently constituted facts or states of affairs, but are instead 'projecting' our own feelings, attitudes, or other affective (as distinct from cognitive) psychological states onto the world. Hume's treatment of moral judgments (according to which morals are "more properly felt than judg'd of") and judgments of causal necessity (as expressive of a felt determination of the mind to infer effect from cause, rather than record some objective necessary connection) are often regarded as classic examples of projectivist theories. As the first example suggests, projectivist theories are often developments of Expressivist (q.v.) accounts of a region of discourse (see also **Quasi-realism**).

Proposition: A proposition is what is said in a speech act, and also the content of a mental act or attitude – for example, a belief or desire state, or a mental act of judgment. A theory that appeals to propositions usually begins by factoring a speech act into content and force. An assertion – for example, a statement that the window is closed – differs in force, but might have the same content as a request that the window be closed. Different theorists differ about what a proposition is, though it is usually assumed to be an abstract object explained in terms of a set of truth-conditions: propositions might be identified either with truth-conditions themselves, or with a structure that represents a recursive procedure for determining a set of truth-conditions.

Propositional attitudes and subdoxastic states: Propositional attitudes are states such as beliefs, desires, hopes, and wishes. The general form of a propositional-attitude ascription can be represented as follows: S j's that P, where "S" stands for the subject to whom the attitude is attributed (John, Jim), "j" for the type of attitude ascribed (belief, desire), and "P" for the informational content of the attitude (that Belhaven is brewed in Dunbar, that Raith Rovers sign Paul Gascoigne). Examples of propositional attitudes are thus John's belief that Belhaven is brewed in Dunbar, or Jim's desire that Raith Rovers sign Paul Gascoigne. Philosophers have attempted to delineate a related but distinct category of mental state, called subdoxastic states: these are like propositional attitudes in so far as they possess informational content, but different from propositional attitudes in so far as they are not ordinarily available to consciousness, and are inferentially insulated from the rest of their possessors' cognitive states: S. Stich, "Beliefs and subdoxastic states," *Philosophy of Science*, 45 (1978); M. Davies, "Tacit knowledge and subdoxastic states," in A. George (ed.), *Reflections on Chomsky* (1989). Philosophers such as Davies, who are impressed by the arguments of Evans – "Semantic theory and tacit knowledge," in S. Holtzmann and C. Leich (eds), *Wittgenstein: To Follow a Rule* (Routledge, 1981) – to the effect that states of tacit knowledge of semantic axioms cannot be propositional attitudes, have attempted to construe states of tacit knowledge as subdoxastic in this sense.

Public language: So called by contrast with Private language (q.v.), a public language is any language in which two or more speakers can communicate with each other. All existing natural languages may be presumed public languages in this sense. The philosophical use and interest of this concept lies in the questions (1) whether the fact that a language is public may not rule out certain accounts of the meaning of its words, and (2) whether being a language at all may not necessitate being a public language. (See Chapter 11, MEANING AND PRIVACY.)

Quantifying in: In (extensions of) first-order logic, a sentence involves *quantifying into* a construction when there is a variable within the scope of the construction bound by an operator without. For example, the sentence

(1) $\forall x(Fx \rightarrow \Box Fx)$

involves quantification into the necessitation construction (the construction which maps a sentence S to '$\Box$S'), as the last 'x' in (1) is in the scope of the construction, and is bound by the initial quantifier, which is not.

One often speaks of quantification in natural language; an example of such is

(2) Each spy believed that the man next to him was a spy

when 'him' is anaphoric on 'each spy.' One may understand such talk in terms of regimentation – a sentence involves quantification into construction c if its regimentation involves quantification into c's regimentation. A first stab at a direct definition might be S involves quantification into c when it contains pronouns (or the like) within c, bound to or anaphoric on expressions which occur outside of c.

Quasi-realism: 'Quasi-Realism' is the name conferred upon a species of anti-realism by its principal proponent, Simon Blackburn. As applied to moral discourse, the quasi-realist maintains, with the Expressivist (q.v.), that to construe moral utterances as descriptive of moral facts or states of affairs is a philosophical error – such utterances are to be understood as expressive of moral feelings or attitudes which we project on to the world, rather than as aimed at recording aspects of an independently constituted moral reality. Blackburn's distinctive aim is to show that the absence of a special realm of non-natural properties or facts to render moral judgments true or false does not – as on an Error theory (q.v.) – entail that such discourse is inherently defective, and that we can quite properly, and consistently with anti-realist scruples and with a Projectivist (q.v.) theory of morality, present our moral sentiments 'in propositional style,' as if they were genuine judgments with truth-conditions.

Radical interpretation: To interpret an item, for example a sound, mark, or gesture, is to assign a meaning to it. When we hear a remark in a familiar language we usually interpret it easily in the light of our knowledge of the language and/or of the person speaking. But radical interpretation is the enterprise of working out the meanings of what we take to be some set of utterances, when we start with no prior semantic knowledge at all, such as information about the structure of the language, the meanings of any words, or the beliefs, interests, and so on of the speakers. It is closely related to the enterprise of Radical translation (q.v.) introduced by Quine; but it differs in that it is to issue not just in claims that a certain sentence (of the native language) and this sentence (of my language) are synonymous, but in actual statements of the meaning of the sentence in the native language. It is thought that reflecting on how we might establish semantic hypotheses from

non-semantic starting points could be a useful philosophical tool in throwing light on the metaphysics and epistemology of meaning.

Radical translation: The process of translating a foreign language in circumstances where no prior knowledge of its syntax, or the etymology and/or likely meaning of any of its expressions, can be assumed, and where no assistance is to be had from bilinguals. Thus the assumption is that the radical translator is restricted, at least initially, to data which exclusively concern the linguistic behavior of the native speakers of the foreign language. It is usually assumed, in addition, that the translator is able from the outset to identify natives' expressions of assent and dissent. It is an important assumption of Quine's philosophy of language that there cannot be more to meanings than can be detected under such constraints. (See also **Radical interpretation** and **Myth of the museum**.)

Referentialism: The term 'Referentialism' is used for the movement in the philosophy of language that rejects the Fregean doctrine that the content of a term (definite description, name, or indexical) is a mode of presentation or an identifying condition. Referentialists such as Marcus, Kaplan, Donnellan, and Kripke claim that, at least for names and indexicals, this is not so. In Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS, a pair of distinctions are made, assigning rather special meanings to some well-worn terms from the philosophy of language, to further clarify this thesis:

Denoting versus naming: An expression denotes when the rules of language assign specific conditions to the expression that an object must meet to be designated by it. We say the expression denotes the object that meets these conditions, if any. Expressions that denote are contrasted with those that name. The rules of language assign specific objects to names, rather than conditions.

Describing versus referring: A term describes if it denotes, and contributes the condition assigned to it by the rules of language, rather than the object denoted, to the content of statements containing it. Describing is contrasted with referring. A term refers if it contributes the object it designates (denotes or names) to the content of statements containing it. Given this terminology, we can say that referentialism is the doctrine that indexicals and names refer rather than describe: indexicals denote and refer, names name and refer.

Relative Identity Thesis: The Relative Identity Thesis is the thesis formulated by Peter Geach in "Identity," *Review of Metaphysics*, 21 (1967), that identity is relative. The thesis has several components. The most ambitious contention Geach puts forward is that absolute identity is inexpressible; a language may contain a predicate expressing indistinguishability by the predicates it contains, but no language can contain a predicate expressing indistinguishability *simpliciter*. Additionally, Geach maintains that identity under a sortal concept (*being the same* A where 'A' is a sortal term) need not entail indistinguishability even by all the predicates contained in a language in which such sortal-relative identity is expressed. Thus objects may be identical under one sortal concept, distinct under another.

Restricted versus unrestricted quantification: The notion of restricted quantification is a significant element in the package of ideas put forward by Peter Geach under the heading of 'relative identity' in "Identity," *Review of Metaphysics*, 21 (1967). It is an important part of his position that sortal terms, in subject position, are names, on a par with proper names; and he takes this to imply that where 'A' is a sortal term 'some A is F' is stronger than 'something is A and F,' and 'every A is F' is weaker than 'everything, if it is an A, is F.' If Geach's thesis that identity under a sortal concept need not be absolute identity is

accepted, the irreducibility of restricted quantification follows. Otherwise there is no argument for it.

Rigidity: Rigidity is a semantic property of expressions. Where e is an expression, let e-c denote the occurrence of e in the context c. e-c is rigid with respect to a class of points of evaluation (e.g., possible worlds, or times) just in case, for some x, the designation of e-c is x, and at every point of evaluation in which x exists, the designation of e-c is x, and e-c has no designation other than x at points of evaluation at which x does not exist. For example, the expression "I" when uttered by Frank is rigid with respect to the class of possible worlds, since in any possible world in which Frank exists, that occurrence of "I" designates Frank, and designates nothing else in worlds in which Frank does not exist. An expression can be said to be rigid just in case the denoting occurrences of it are.

The concept of rigidity has proved to be a useful tool in gaining an understanding of the complex semantical interactions between intensional operators and certain classes of expressions, such as names, indexicals, and pronouns. Rigidity is also important in the model theory of quantified intensional logic, where it is used as a restriction on variables.

S4 principle: The distinctive theorem schema of C. I. Lewis and C. H. Langford's system S4 of modal logic can informally be stated thus: if it is necessary that P, then it is necessary that it is necessary that P. This is expressed by the formula $\Box p \rightarrow \Box\Box p$. Readings of the symbol $\Box$ other than 'it is necessary that …' give what may also be described as S4 principles. For example, if $\Box$ is read as 'it is clear that …,' the result is the principle that if it is clear that P, then it is clear that it is clear that P, a principle to which Higher-order vagueness (q.v.) may provide counter-examples. In possible-worlds semantics the S4 principle corresponds to the condition that if a world z is possible from the standpoint of a world y, and y is possible from the standpoint of a world x, then z is possible from the standpoint of x, that is, relative possibility is transitive.

S5 principle: The distinctive theorem schema of C. I. Lewis and C. H. Langford's system S5 of modal logic can informally be stated thus: if it is possible that P, then it is necessary that it is possible that P. An equivalent schema is: if it is not necessary that P, then it is necessary that it is not necessary that P. This is expressed by the formula $\neg\Box p \rightarrow \Box\neg\Box p$. Readings of the symbol $\Box$ other than 'it is necessary that …' give what may also be described as S5 principles. For example, if $\Box$ is read as 'it is clear that …,' the result is the principle that if it is not clear that P, then it is clear that it is not clear that P, a principle to which Higher-order vagueness (q.v.) may provide counter-examples. In possible-worlds semantics, the S5 principle corresponds to the condition that if worlds y and z are possible from the standpoint of a world x, then z is possible from the standpoint of y. Given that relative possibility is reflexive, this is equivalent to the condition that it should also be both symmetric and transitive.

Second-order languages: see **First-order languages**

Semantic conception of vagueness: What distinguishes this view from the Ontological conception of vagueness (q.v.) is the claim that all vagueness arises in thought and talk, not in what is thought and talked about (in so far as it is not itself thought and talk). What distinguishes the semantic conception from the Epistemic conception (q.v.) is the claim that vagueness results in the failure to make a statement that is either true or false (and not both) in borderline cases. There is either no truth-value at all, or one of a non-standard kind, for example 'neutral,' 'true to degree 0.7,' or 'both true and false.' Such behavior may be attributed either to gaps and defects in the meanings of vague expressions, or to positive meanings of an alternative kind, for instance as given by prototypical examples

uncircumscribed by boundaries. Many different technical treatments of vagueness are consistent with the semantic conception. (See **Degree of truth**, **Supervaluation**.)

Semantic creativity and learnability: Speakers of a natural language display semantic creativity in so far as they are able to understand novel utterances, that is, utterances of sentences which they have never before encountered. For example, your understanding of "Napoleon's grandfather wore purple pyjamas" is (almost certainly) a manifestation of semantic creativity. A related idea is that of the learnability of natural languages: a natural language is said to be learnable when a speaker needs only explicit training with, or exposure to, a small part of the language, in order to secure competence with a larger and more extensive part.

Semantic externalism: Semantic externalism is the view that the propositional contents of at least some of the sentences uttered by a speaker fail to supervene on the intrinsic features of the speaker. The opposing view, semantic internalism, claims that the propositional contents *do* so supervene. Externalist views are typically argued for using thought experiments featuring internal duplicates, that is, speakers who by hypothesis share all intrinsic features, but whose utterances putatively possess different propositional contents. Sometimes externalist views are limited to some classes of expression such as natural kind terms (see **Natural kind externalism**), while other externalist views are more wide-ranging, appealing for example to the role of linguistic communities in determining content (see **Social externalism**).

Semantic holism: Semantic holism is the view that meaningful items (1) must necessarily occur as part of some whole, that is, a large set of such items, and (2) are such that the meaning of each item is somehow bound up with, constrains, and is constrained by, the meanings of the other items. It comes in various different versions (often not distinguished as sharply as would be useful) depending on whether the items are taken to be sentences and the whole the language of which they are elements, or whether the items are sentences and the whole a theory they jointly compose, or whether the items are thoughts and the whole a mind of which they are the contents.

Semantic internalism: see **Semantic externalism**

Semantic irrealism: Semantic irrealism is the thesis that there are no semantic facts and hence, in particular, no facts about what any expression means. Outrageous and even paradoxical as it may seem, this thesis has been seriously advanced and defended: most famously, perhaps, Saul Kripke takes it to be established by the skeptical argument he extracts from Wittgenstein's remarks on rule-following; and W. V. O. Quine draws a similar conclusion from his arguments for the indeterminacy of translation. Note that the semantic irrealist should not be seen as claiming that all expressions are meaningless, since the fact that an expression lacks meaning would be just as much a semantic fact as would the fact that it means such-and-such. Nor is the semantic irrealist necessarily committed to denying that semantic sentences have any proper use – though Quine certainly denies that they can have any part in serious science. The essential irrealist claim is that such sentences – in contrast with, say, sentences belonging to physics, for example, or geography – do not have meaning by being associated with truth-conditions. This leaves room to hold that semantic sentences have meaning in some other way, just as denying that there are moral facts leaves room to hold that moral utterances have meaning in some non-truth-conditional way. (See Chapter 26, INDETERMINACY OF TRANSLATION, for discussion of Quine's position, and Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING, for discussion of Kripke's.)

Semantic naturalism: Semantic naturalism is the view that semantic and intentional properties and relations, especially the properties of referring to something and of having such-and-such a truth-condition, are part of the natural order. They are instantiated in virtue of the instantiation of natural properties and relations. Natural properties and relations are ones that are reducible to (or realized by) properties and relations expressible in the vocabulary of the natural sciences (physics, chemistry, and biology) and the causal and nomological facts involving them. Most contemporary semantic naturalists think that the semantic properties of natural-language expressions can be explained in terms of the semantic properties of the mental states of users of the language, and that the semantic properties of mental states can be explained in terms of causal relations between them and various extra-mental items.

Semantic physicalism: see **Semantic naturalism**

Semantic value: A semantic value for an expression is an object that a semantic theory that interprets the language assigns to the expression. The term is intended to be abstract and neutral, leaving open what kind of objects a semantic theory uses to interpret its expressions. Semantic values for whole sentences may include the propositional content or thought expressed by the sentence, or a function that determines content as a function of context. The semantic values of words and other constituents of sentences will be whatever, according to the semantics, they must be to determine the semantic values of the expressions that contain them as constituents. In extensional theories, semantic values will be extensions.

Sense: Sense is that with which meaningful expressions are invested, and that in virtue of which they make reference to non-linguistic items (or, speaking more generally, have semantic values). Frege introduced the idea of the sense of an expression to be correlative with the idea of a speaker's understanding of the expression (see Chapter 2, MEANING AND TRUTH-CONDITIONS, especially §2). The sense of an expression is, then, the expression's *contribution* to the sense of any larger unit which can be used to say something true or false. In the limiting case of a sentence, the sense is the truth-condition. Or better (postponing commitment to senses as entities): to know what is the sense of a sentence is to know under what conditions the sentence has the semantic value of truth. Frege calls the sense of a sentence, which is very much the same thing as many philosophers have intended by "proposition," a *thought*. (The psychologistic overtones of the English word must be carefully kept out here. The thought is something public and non-psychological.)

Another special case of sense is that of the sense of a singular term. (Frege sets out from this case in "On sense and reference" in P. Geach and M. Black (eds), *Philosophical Writings of Gottlob Frege*, Blackwell, 1952.) A singular term has its sense by standing for – or presenting – its reference, which is an object. Different singular terms may present one and the same object by different "modes of presentation." Wherever this is the case, the singular terms in question will make different contributions to the overall sense of the various sentences in which they figure. To grasp the sense of a proper name fully and correctly, then, is to know which object the name stands for, and to know this in virtue of being party to a particular way of thinking of this object. Different ways may call into being different proper names with different senses, which may contribute to different thoughts about one and the same thing, and may thereby contribute to an explanation of the possibility of informative statements of identity. This feature of the Fregean scheme has attracted the skepticism of those who see problems in the individuation

of different ways of thinking of an item, and of those who see naming as irreducible to describing, and believe that different ways of thinking of a thing collapse into different descriptions of the thing.

Singular proposition: Suppose that assertion and belief are relations; call their objects *propositions*. Broadly Fregean views of propositions see them composed and individuated in terms of "ways of thinking" of objects and properties; broadly Russellian and Millian accounts posit *singular propositions*, whose identity is a function simply of the objects and properties they concern. A Fregean representation of the proposition that Fichte weeps, pairs a way F of thinking of Fichte and a way W of thinking of the set of weepers; the singular proposition may be represented by pairing Fichte with the property weeping.

As propositions determine truth-conditions, one difference between the views is truth-conditional. The Fregean proposition's truth presumably turns on the object, which F presents, having the property W presents; the singular proposition is true just in case Fichte weeps. If F can present something other than what it in fact does (or fail to present Fichte, though he existed), the propositions have different truth-conditions. A second difference is this: if 'Fichte weeps' expresses the Fregean proposition, presumably one must think of Fichte in way F to think that Fichte weeps; if it expresses a singular proposition, this is not so.

Social externalism: A version of semantic externalism, according to which the contents of many (or possibly all) linguistic expressions used by a speaker are partly determined by features concerning the other members of the speaker's linguistic community. The best known argument for social externalism is Tyler Burge's *arthritis* thought experiment. (See **Semantic externalism**.)

Sortal concepts: The notion of a sortal concept is the notion of a concept which conveys a criterion of identity and thus determines a type of object for which it makes sense to ask whether objects of the type are the same or different. Although the notion of a sortal concept can be illustrated by standard examples – 'man,' 'gold,' 'number,' and 'direction' stand for sortal concepts – it is controversial how it is to be explained. Geach's view, in "Identity," *Review of Metaphysics*, 21 (1967), which is part of the package he offers under the title of 'relative identity,' is that predicates expressing sortal concepts like 'is a man' are semantically derivative from relational predicates expressing equivalence relations, like 'is the same man as': 'is a man' must be understood as 'is the same man as something' just as 'is a brother' must be understood as 'is a brother of someone.' This view is opposed by Michael Dummett in 'Does quantification involve identity?' in H. A. Lewis (ed.), *Peter Geach: Philosophical Encounters* (Kluwer, 1991), who argues that in the case of many sortal predicates, such as 'is a number,' they are rather to be understood as derivative from functors, for example 'the number of ….' Whatever the correct way to understand the notion of a sortal concept, it is generally accepted that both Mass terms (q.v.), which provide a criterion of identity but no principle for counting, and count nouns, which provide both, are sortal terms.

Speaker-meaning: A statement of the form: 'By uttering x, U meant such-and-such' aims to report what an agent meant$_{nn}$ (see **Natural and non-natural meaning**) by doing something (perhaps uttering certain words) on a particular occasion. This will be an instance of speakers' occasion-meaning$_{nn}$. By contrast, 'U means such-and-such by x' would ordinarily be used to say something about what U typically means$_{nn}$ by x, whenever she utters x – this will be an example of speakers' 'timeless' meaning, in Grice's terminology.

Speakers' occasion-meaning is the analysandum concept in the Gricean analysis of meaning. It is what speakers mean that is analyzed in the first instance. In Grice's approach, one builds up to the concept of linguistic meaning from that of speaker-meaning. The priority of speaker-meaning in this approach is evidence that the concept of meaning that Grice is concerned to analyze is somewhat *wider* than linguistic meaning.

Stimulus synonymy: Quine's behavioristic, non-normative surrogate for synonymy as more ordinarily conceived. A pair of sentences are stimulus-synonymous just in case the same stimuli as provoke assent to (or dissent from) the one provoke assent to (respectively, dissent from) the other; a pair of sub-sentential expressions are stimulus-synonymous just in case any sentence containing an occurrence of the one is stimulus-synonymous with the result of replacing that occurrence by one of the other.

Strict implication and equivalence: A statement A strictly implies a statement B if and only if it is necessarily true that if A then B (i.e., if and only if $\Box (A \rightarrow B)$, where '$\rightarrow$' is the material or truth-functional conditional). A and B are strictly equivalent if and only if they strictly imply each other (equivalently, if and only if $\Box (A \leftrightarrow B)$, where '$\leftrightarrow$' is the material or truth-functional biconditional).

Subject-matter: What a sentence or piece of discourse is about. This relates not only to the objects mentioned but also to the facts that might bear upon the truth or falsity of the sentence or discourse. Subject-matter in this latter sense might be represented by an equivalence relation on worlds or by a suitable closed set of candidate truthmakers; and the notion of subject-matter has played a significant role in the development of truthmaker semantics. (See Chapter 22, TRUTHMAKER SEMANTICS.)

Substitutional quantification: To understand "$(\exists x)Fx$" substitutionally is to understand it as true iff there is some name (say *n*) which can be substituted for the "x" in "Fx" such that the resulting sentence F*n* is true. This contrasts with the more common objectual understanding of "$(\exists x)Fx$," whereby it is true iff there is some object which is F. If the only object or objects which are F lack names, "$(\exists x)Fx$" will be true on the objectual reading but false on the substitutional. This difference will disappear if one extends the substitutional account to include new names, arguing perhaps that any object can be *given* a name. That, however, looks suspiciously like reducing substitutional quantification to objectual after all. The substitutional account may seem to help with "Whatever the Pope says is true": for any value of s, if the Pope says "s," then s. But it must not be allowed to disguise the problem of whether "s" is functioning in two different ways here. A decision to read the quantifier substitutionally does nothing by itself to explain the relation between the two occurrences.

Substitutivity, principle of: see **Leibniz's Law**

Superassertibility: This is a notion proposed originally by Crispin Wright as an anti-realistically acceptable replacement for the realist notion of Evidence-transcendent truth (q.v.). It is superassertible that P if and only if there is, or can be, warrant to assert that P, and some warrant to assert that P would survive arbitrarily close scrutiny of its credentials and arbitrarily extensive increments to, or other improvements upon, our state of information.

Supervaluation: Most generally, a supervaluation is a semantic property defined by quantification over valuations. The idea seems first to have been used in connection with philosophy of science, the name in connection with Liar paradoxes: for the history, see Williamson, *Vagueness* (Routledge, 1994), §5.2. In connection with vagueness, the valuations are classical assignments of sets to predicates ("sharpenings"); admissible valuations

meet various constraints designed to ensure that between them they represent the meaning of the predicate, even if it is vague. (So, e.g., an admissible valuation does not assign to the extension of the predicate something to which, intuitively, it definitely does not apply.) Truth (or "supertruth") is defined as truth-upon-all-admissible-valuations, and falsehood as falsehood-upon-all-admissible-valuations. This leaves room for the possibility, supposedly indicative of vagueness, of sentences which are neither true nor false.

Supervenience: The basic idea is that one range of facts supervenes upon another (the base or subvening facts) iff there could be no differences among the supervening facts without a difference in the subvening ones. When it is said that meaning supervenes upon use, the idea is that there could not be two communities who count as speakers of different languages unless there were differences in how each used its language. In serious discussion, some account must at once be given of "use," else the claim is trivial. (For example, it is not very controversial that the meaning of a declarative sentence supervenes on what it can be correctly used to say.) The slogan is often associated, historically, with behaviorist views in semantics: those which regard use as capable of being specified wholly behavioristically.

Tacit knowledge: Philosophers of language have invoked this notion in attempts to explain the Semantic creativity and learnability (q.v.) of natural language. The main idea is that if speakers of a language can be credited with tacit knowledge of a compositional semantic theory for their language (see **Compositionality**), we will have an explanation of semantic creativity (since the theory the speaker tacitly knows has the resources to generate a meaning-specifying theorem for the sentence uttered in the novel context) and an explanation of learnability (since the theory tacitly known can generate meaning-specifying theorems for a wide range of sentences on the basis of a narrower range of semantic axioms). Philosophers disagree as to the nature of states of tacit knowledge; in particular, they disagree as to whether they should be construed as *bona fide* propositional attitudes – see M. Dummett, "What is a theory of meaning? (II)," in G. Evans and J. McDowell (eds), *Truth and Meaning* (Oxford University Press, 1976) – or as mere dispositional states – see G. Evans, "Semantic theory and tacit knowledge," in S. Holtzmann and C. Leich (eds), *Wittgenstein: To Follow a Rule* (Routledge, 1981) – or as subdoxastic information-containing states – see M. Davies, "Tacit knowledge and the structure of thought and language," in C. Travis (ed.), *Meaning and Interpretation* (Blackwell, 1986).

Theory of meaning: As it appears in discussions by philosophers of language, this phrase may be taken in at least two ways. In the more general of the two senses, it denotes a theory dealing with language in general, which attempts to analyze and elucidate the concept of meaning. In this sense, it can be applied to attempted analyses or conceptual elucidations of what it is for an item to have meaning, as for example in Grice's linkage of linguistic meaning with speakers' intentions or in verificationist-style accounts which insist that what we can think, and hence mean, is importantly constrained by what we can know (see **Epistemic conception of meaning**). In the more specific sense it denotes a theory dealing with a particular language, which generates a theorem specifying the meaning of each well-formed declarative sentence of that language. Philosophers disagree about the nature of these theorems (often called "meaning-specifying theorems"): some take them to be statements of sentences' truth-conditions (Davidson), while others take them to be statements of sentences' conditions of warranted assertibility (Dummett).

Theory of truth, Tarski-style: The phrase 'a Tarski-style theory of truth' is used to describe an assignment of semantic properties to words and constructions which enables us,

given a specification of a sentence as built from certain words by certain constructions, to work out a biconditional stating truth-conditions of that sentence. Tarski himself insisted, for his purposes of giving the so-called semantic theory of truth, that the theory should issue in statements of truth-conditions in which the sentence which gives the truth-conditions has the same meaning as the sentence whose truth-conditions are given. (In the case where the metalanguage includes the object language, this results in the demand that the output of the theory should be theorems like the famous '"Snow is white" is true iff snow is white.') But the idea of a 'Tarski-style theory of truth,' as invoked, for example, by Davidson in the context of his project of Radical interpretation (q.v.), abandons that requirement; the idea is to use the formal apparatus devised by Tarski (e.g., of connectives or quantifiers) for a different purpose. (See also **Compositionality**.)

Things: There is a very weak sense of the term 'thing' according to which it is trivially true that *everything* is a thing. In this sense, any item which a system of Ontology (q.v.) acknowledges as existing at all may be accounted a 'thing' in that ontology. But for most purposes the term 'thing,' or 'object,' is used in a more robust sense, in which its application is restricted at least to the class of entities possessing well-defined identity-conditions, and hence to Kinds (q.v.) of entity for which Criteria of identity (q.v.) may be stated. Sometimes, however, 'thing' is used in an even more narrow sense than this, to refer to concrete, physical occupants of space which persist through time (otherwise called 'continuants') – a sense which excludes both Abstract objects (q.v.) and such physical entities as events. Clearly, if confusion is to be avoided in the use of the word 'thing,' care must be taken to indicate which of these senses is in operation.

Thought: see **Sense**

Timeless meaning: Timeless meaning is the meaning our utterances have when they are not tied to a particular occasion of utterance. Linguistic meaning is an instance of timeless meaning. If one takes it that linguistic meaning is necessarily structured (as philosophers do), then it is possible to have timeless meaning that is not linguistic. A system of communication based on (unstructured) gestures would be an example. Most – though not all – Griceans propose to build up to the notion of timeless meaning by adding the concept of convention to the analysis of Speaker-meaning (q.v.).

Tokens: Consider this list: philosophy, art, history, philosophy. Are there three or four words on the list? Three types appear, but four tokens; there are two tokens of the type "philosophy." Sometimes "token" is used for a particular act of uttering an expression of a given type, but more often it is used for the object that is produced by such utterances: the particular ink marks on a page, or the burst of sound that travels through the air. In this volume (see Chapter 38, THE SEMANTICS AND PRAGMATICS OF INDEXICALS) "token" is used in the second way, and "utterance" is reserved for acts. In this usage, a token is an effect of an utterance.

Travis cases: A Travis case (named after Charles Travis) is an imagined (or real) scenario where tokens of a linguistic type may have varying truth-values over a fixed context, even though the type is unambiguous and contains no indexical items. The putative consequence of the coherence of such scenarios is that meaning does not determine truth-value even with all relevant parameters and indexical factors fixed. Consider an utterance of 'The leaves are green' in a scenario where the relevant leaves are painted green. The crucial intuition is that the utterance may be read as true or false depending on a divergence between speakers' non-linguistic interests or purposes. A photographer, say, may happily take the sentence to be true, for she is only concerned with the surface

properties of the leaves, whereas a botanist is concerned with the natural state of the leaves, not merely how the leaves appear. Many of the disputes concerning such cases, engaged in by both linguists and philosophers, pertain to their significance for traditional compositional truth-conditional semantics, for such cases appear to show that fixing the meanings of the words in a sentence and the semantic significance of the syntax of the sentence is not sufficient to fix truth-conditions for the sentence, even relative to context/circumstances.

Truth-bearers: There has been much dispute over what the (primary) bearers of truth are. Candidates have included sentences, statements, judgments, propositions, thoughts, and beliefs. The dispute has been partly verbal. A sentence can be true, but evidently what is true is the sentence *as* used on a particular occasion, to convey a particular message; in other words, to make a particular statement or to express a particular proposition. Hostility to statements and propositions has often arisen from thinking they must be entities of a peculiar metaphysical kind. Some defenders of propositions have indeed thought of them like that, but without such commitment one can legitimize talk of propositions if one can specify the conditions under which two utterances express the same proposition. Following Quine, some have maintained this cannot be done – a serious objection, if correct, but at least counter-intuitive. The objection to taking thoughts, judgments, or beliefs to be primary truth-bearers is that they are subjective mental occurrences. But again, while beliefs (etc.) are held by individuals, there seems a good sense in which two people can be said to share the same belief – which they might voice by uttering sentences expressing the same proposition.

Truth-condition: see **Sense**

Truthmaker: An object in the world (such as a fact or state of affairs) that makes a sentence true. The truthmaker can be: exact, wholly relevant to the sentence it makes true; inexact, partially relevant to the sentence; or loose, merely standing in a modal relationship to the truth of the sentence. Truthmakers are of importance, both in metaphysics as a guide to the structure of the world and in semantics as a guide to meaning; and the exact form of truthmaking has recently been thought to be especially important in this regard.

Underdetermination of empirical theory by data: The contention that any body of empirical evidence for a particular theory will be compatible with alternatives to that theory, and hence cannot constrain the selection of any particular one among a range of alternative empirical theories. The thesis is uncontentious if one restricts attention to finite accumulations of evidence. In Quine, however, it is generalized to cover *all possible* empirical data. A stronger version yet contends that even all possible data *plus* best scientific methodology – proper canons of simplicity, economy, …, integration, and so on in theory construction – never determine one particular empirical theory as uniquely best. The Underdetermination Thesis is the main premise for Quine's principal argument for the Indeterminacy of translation (q.v.).

Universals: Universals are the (supposed) referents of general terms, such as 'red,' 'car,' and 'planet,' conceived as entities distinct from any of the Particulars (q.v.) which instantiate them. As such, they are Abstract objects (q.v.). Whether such entities really exist has long been a matter for heated debate, those denying their existence usually being called 'nominalists' and their opponents 'realists.' There are two main schools of realism: Aristotelian or 'immanent' realism, which holds that universals exist 'in' particulars, and accordingly cannot exist if uninstantiated; and Platonic or 'transcendent' realism, which holds that universals exist 'separately' from particulars, and consequently may exist even if they have no instances.

Use theory of meaning: Use theories of meaning are sometimes referred to as communication or pragmatic theories. Grice's account of meaning is an example of a use theory. Such theories of meaning have their roots in the work of J. L. Austin (the idea that meaning is to be associated with use is also a theme in the later writings of Wittgenstein). Such approaches to meaning connect the concept primarily with the actions of speakers. The approach looks at the phenomenon of meaning quite widely, taking linguistic meaning to be a particular instance of the phenomenon. As a result, structure is accorded less emphasis.

Verificationism: The verifiability criterion of meaningfulness is the claim that a meaningful sentence is one which is capable of being verified or falsified. Even those who would generally find no inclination to hold such a view find themselves disposed to think that, if our vague predicates really do draw sharp boundaries, it would have to be possible for us to tell where these boundaries fall. Such theorists would do well to justify the specific view about vagueness on some basis other than the general verificationist claim, for this latter is now rarely thought defensible, at least in the crude form given here.

A verificationist theory of meaning holds, in general terms, that understanding a sentence consists in grasping what information states would verify it. An information state verifies a sentence just if a person in that state is warranted in asserting it. Strict forms of verification further require that the verifying information state should be indefeasible.

A verificationist view of truth holds that truth is verifiability. A sentence is true if and only if it is verifiable, that is, if and only if there is evidence warranting its assertion.

Wide cosmological role: One of several features or marks proposed by Crispin Wright, in virtue of which realism concerning the subject-matter of a given discourse whose characteristic statements qualify for at least Minimal truth (q.v.) may be maintained. A discourse exhibits wide cosmological role if the facts recorded in its characteristic claims have a role to play in explanations of further facts of other kinds, beyond facts about our beliefs and other attitudes, and can figure in such explanations other than as objects of those attitudes. It might be contended that while moral, or modal, beliefs, for example, are apt to figure in explanations of our actions, desires, and other beliefs, moral or modal facts themselves exert no influence on other goings-on; in contrast, facts about the primary qualities of bodies, for example, exert causal influence in the world at large. To the extent that this is so, we might think that it justifies a kind of realism about facts of the latter kind which is unwarranted in regard to the former. (See **Cognitive command**, **Euthyphro contrast**.)

Index

Page numbers in *italics* indicate figures; page numbers in **bold** indicate tables.

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.