The Inscrutability of Reference

John Robert Gareth Williams

Thesis in candidature for the degree of Ph.D. in Philosophy,
University of St Andrews

As submitted Friday, 19th August 2005
Accepted 16th November 2005

Abstract

The metaphysics of representation poses questions such as: in virtue of what does a sentence, picture, or mental state represent that the world is a certain way? In the first instance, I have focused on the semantic properties of language: for example, what is it for a name such as ‘London’ to refer to something?

Interpretationism concerning what it is for linguistic expressions to have meaning, says that constitutively, semantic facts are fixed by best semantic theory. As here developed, it promises to give a reductive, universal and non-revisionary account of the nature of linguistic representation.

Interpretationism in general, however, is threatened by severe internal tension, due to arguments for *radical inscrutability*. These contend that, given the interpretationist setting, there can be *no fact of the matter* what object an individual word refers to: for example, that there is no fact of the matter as to whether “London” refers to London or to Sydney.

A series of challenges emerge, forming the basis for this thesis.

1. What sort of properties is the interpretationist trying to reduce, and what kind of reductive story is she offering?

2. How are inscrutability *theses* best formulated? Are arguments for inscrutability effective in their own terms? What kinds of inscrutability arise?

3. Is endorsing *radical inscrutability* a stable position?

4. Are there theoretical virtues—such as simplicity—that can be appealed to in discrediting the rival (empirically equivalent) theories that underpin inscrutability arguments?

In addressing these questions, I concentrate on diagnosing the source of inscrutability, mapping the space of ways of resisting the arguments for radical inscrutability, and examining the challenges faced in developing a principled account of linguistic content that avoids radical inscrutability. The effect is not to close down the original puzzles, but rather to sharpen them into a set of new and deeper challenges.

Chapter list

Contents

Acknowledgements

I prepared this work while a PhD student within Arché, the AHRC centre for the philosophy of logic, language, mathematics and mind, a research unit within the Departments of Philosophy in the University of St Andrews. I was privileged to have the opportunity of working in this intensive research culture. I gained much from attending and presenting material at seminars in all three Arché research projects: on the Foundations of Classical Mathematics, on Modality and on Vagueness, and also from material presented and discussed in the Arché postgraduate discussion group. The informal discussion of research that the setting allows has been invaluable.

There are many individuals that I wish to thank; in compiling acknowledgements, one runs the danger of offending those who one forgets to list. Since I wish to mention people by name, it is almost inevitable this situation will arise. Apologies in advance, therefore, and let me take this opportunity to express general gratitude to all those who have helped me in my researches, and made life such fun.

I wish to thank my colleagues in Arché, past and present, for philosophy, advice, and support: Elizabeth Barnes, Eline Busck, Aislinn Batstone, Ross Cameron, Roy Cook, Richard Dietz, Philip Ebert, Bob Hale, Aviv Hoffman, Carrie Jenkins, Patrick Greenough, Dan López de Sa, Fraser MacBride, Darren McDonald, Aidan McGlyn, Sebastiano Moruzzi, Josh Parsons, Nikolaj Pedersen, Graham Priest, Agustín Rayo, Silvia Rescigno, Sonia Roca, Marcus Rossberg, Anna Sherratt, Stewart Shapiro, Chiara Tabet and Elia Zardini.

Within the Philosophy department more generally, I would like to thank: Peter Clark, Joe Diekemper, Jamie Dow, Katherine Hawley, Iwao Hirose, Kent Hurtig, Sune Holm, Janet Kirk, Brian McElwee, Daniel Nolan, Simon Prosser, Steve Read, Sujan Riyadh, Simon Robertson, Enzo Rossi, Marcus Schlosser and Michael Weh. I would like to thank an institution: the St Andrews postgraduate 'friday' seminar.

Arché visitors and associate fellows have been a constant source of stimulation: in particular, I am indebted to discussions with JC Beall, Bill Brewer, John Divers, Cian Dorr, Dorothy Edgington, Matti Eklund, Kit Fine, John Hawthorne, Richard Heck, Frank Jackson, Michael Lynch, Vann McGee, Andy McGonigal, Sven Rosenkranz, Bob Stalnaker, William Stirton, Gabriel Uzquiano, Brian Weatherson and Timothy Williamson.

I would also like to thank Brad Amour Garb, Corine Besson, Wylie Breckenridge, Neil Cooper, Imogen Dickie, Daniel Isaacson, Stephan Leuenberger, Rory Madden, James Moratu, Richard Price, Daniel Rothschild, Wolfgang Schwarz, Richard Smith and Bruno Whittle.

Early versions of chapters have been presented in a number of settings, and I have received valuable feedback from many people in these and other contexts. I have tried to acknowledge specific points that people have brought to my attention in the main text, but more general debts are listed below:

Chapter 1. Particular thanks to Dan López de Sa, who commented on a late draft of this chapter. *Chapter 2.* An early version of this chapter was presented to the Arché mathematics seminar. Particular thanks to Agustín Rayo, who commented on a late draft of this chapter. *Chapter 3.* An early version of this chapter was presented to the Arché vagueness seminar. Thanks are in particular due to JC Beall, Stewart Shapiro, Timothy Williamson and Elia Zardini for discussion of the material on global validity and revisionism. *Chapter 4*. Some of the material in this chapter was presented to the Metaphysics Reading Group, St Andrews. Thanks in particular to Wolfgang Schwarz for discussion and criticism of an early version of this chapter, and to Wylie Breckenridge and Daniel Rothschild for correspondence on specific points. *Chapter 5* Thanks to Achille Varzi for email correspondence. *Chapter 6*. Some of the material included in this chapter was presented under the title "Complex numbers and reference" at an Arché mathematics workshop. *Chapter 7*. Material in this chapter was presented to the Arché Mathematics Project Seminar, St Andrews, and at the Ockham Society, Oxford. I benefitted in particular from discussion with Ofra Magidor, Richard Price and Bruno Whittle. Special thanks go to Carrie Jenkins and Simon Prosser for comments on late versions of this chapter. *Chapter 8*. Material drawn from this chapter was presented at the St Andrews Graduate Seminar, during a short presentation at Leeds, and at the Fourth Barcelona Workshop on the Theory of Reference. Thanks to all present for much valuable discussion and advice. Particular thanks to Elizabeth Barnes, Brian McElwee, Philip Ebert, Chris Gauker, Stephen Leuenberger, Andy McGonigal, Joe Melia, Marcus Rossberg, Daniel Rothschild and Bob Stalnaker for valuable criticism and advice.

My research was funded from 2000-2004 by the Arts and Humanities Research Board, first while on a masters award, during my time in Oxford, and later on a doctoral award in St Andrews. I worked as a half-time Research Assistant at Arché in 2004-5, and am extremely grateful for a scholarship from the departments of philosophy at St Andrews that provided financial assistance during this period.

Two people in particular have improved this thesis at innumerable points: my supervisor, Crispin Wright; and my officemate and *haecceity*-sharer, Ross Cameron. Special thanks to them for many many valuable discussions.

Phyllis Williams, who proof-read the penultimate version of this thesis improved the text immeasurably.

My final thanks go to those who have offered friendship and support throughout. In addition to those already mentioned: Caroline Ayre, Wendy and Robert Bailey, Bec James, Daniel Gifford, Frieda Klotz, Zoë Green, Dave Hughes, Aideen Lee, Graeme Napier, Tim Payne, Jo Ward, and other friends from Wales, Christ Church, Wolfson and beyond. Lastly, to Paula Curnow; my parents Phyllis and Mark Williams; my sisters Jen Williams and Annie Hughes; my grandparents Pat and John Simpson; and the memory of Barbara and Howard Williams; deepest affection and gratitude.

Anticipation

My theme is the inscrutability of reference: the idea that, in a range of cases, there may be no fact of the matter about what it is our words refer to. The thesis is divided into four parts:

1. The framework;
2. Arguing for inscrutability;
3. Against radical inscrutability;
4. Avoiding radical inscrutability.

In this introduction, I shall sketch the contents of each part, and of the chapters within them.

Part I: The framework

Some think that arguments for the inscrutability of reference are perfectly general, or generalizable.[1] I do not think this is the case: I regard them rather as local issues arising within particular philosophical treatments of the representational content of language. Part I describes and develops "interpretationism" as a reductive account of linguistic content in which inscrutability puzzles do arise.

Chapter 1 "The semantic project" describes the general landscape in which the thesis operates, and lays out terminology. It contains three sections. First, I discuss the motivation for attributing semantic properties to words. Second, I make the Quinean distinction between the theory of reference and the theory of meaning, and locate versions of Davidsonian truth-theoretic and Montagovian intensional semantics within this. I discuss possible relations between semantics and an account of the cognitive significance of sentences.

Third, I characterize the *metasemantic* project: to give an account of the nature of the facts apparently uncovered by first-order semantics. I accept Fodor and Field's challenge to *reduce* the intentional to the non-intentional. Finally, I explain how 'interpretationist' metasemantics can be seen as answering this challenge, and briefly describe three versions: Lewis' convention-based approach, so-called 'global descriptivism', and Davidsonian radical interpretation.[2]

Chapter 2 "Reductive Paraphrase" offers an account of interpretationism as a metaphysics of meaning. It construes it as a *reductive paraphrase*—systematically pairing semantic vocabulary ('refers' 'is true') with non-semantic surrogates. The main body of the chapter develops the conception of paraphrase in play. I argue for a Quinean conception of the significance of paraphrase, as constrained by the philosophical purposes at hand, rather than as uncovering the semantic or pragmatic content of the problematic discourse. I set out a form of paraphrase common to a number of philosophical 'fictionalisms', and look in detail at a general problem for such views: the puzzle of 'incomplete fictions'.[3]

[1] Putnam's notorious 'just more theory' argument attempts this. It is briefly discussed in Chapter 8, §8.1 below.

[2] See Davidson (1973), Lewis (1975) and Lewis (1984), respectively. Further references are given in the chapter.

[3] The puzzle is formulated, for the case of modal fictionalism, in Rosen (1990).

I outline how one can view interpretationist metasemantic theories as allowing a 'theory-shadowing' (i.e. broadly fictionalist) paraphrase, reducing the semantic to non-semantic facts. I defend the account in detail against charges that the apparatus induces circularity.

Part II: Arguing for inscrutability

I characterized inscrutability of reference as the thesis that there is 'no fact of the matter' about what our words refer to. The second part of the thesis examines what 'no fact of the matter' comes to within the interpretationist framework, and develops several arguments for disturbing—even untenable—forms of inscrutability within the interpretationist setting.

The first section of Chapter 3, "Inscrutability theses", describes how arguments for inscrutability can be based on the interpretationist metasemantic proposal. The constraints on the 'meaning-fixing' semantic theory is just that it account for a set of data pairing sentences with truth-values or propositions, so if a range of semantic theories all accomplish this, there is *prima facie* no choosing between them. If there is no fact of the matter about which semantic theory fixes meaning, there will be no fact of the matter which of the various (incompatible) claims they make hold good.

The main question this chapter addresses is: how are we to formulate the *claim* that reference is inscrutable, both as theorists, and as speakers of the language in question? Two proposals are described and compared. The first is a version of 'supervaluationism'; the second rests entirely on the way of paraphrasing semantic facts put forward in Chapter 2. The proposals are clearly distinct, but neither seems to have a clear advantage over the other. In this context, I defend supervaluationism against the charge that it engenders departures from classical logic.[4] In Appendix A I argue that even if we treat consequence 'globally', a proper and well-motivated supervaluational framework based on a classical language will accord with classical logic in its entirety.

Chapter 4: "Gavagai again" focuses on Quine's most famous argument for the inscrutability of reference: the 'argument from below' from "Word and Object" (Quine, 1960). I distinguish two types of inscrutability promoted by Quine's discussion: inscrutability over the analysis of predication (e.g. over cashing-out predication in terms of sets and set-membership, universals and instantiation, mereological sums and part-whole, or simply by utilizing a primitive notion of 'application') and inscrutability over the division of reference (e.g. whether 'rabbit' applies to rabbits, instantaneous rabbit-stages, or undetached rabbit parts, for example). *Prima facie* the two issues are orthogonal, and I focus on the latter.

I outline certain technical objections to the Quinean argument for 'division inscrutability'.[5] My response is to set out three systematic semantic theories of fragments of English, that divide reference along the three lines that Quine suggests. These are drawn initially from the literature on persistence: worm theory, stage theory, and a generalization of the 'counterpart theoretic' ideas of Lewis (1968); Sider (1996a) to spatial extension.[6] I then examine how the technical objections play out. My conclusion is that, at least in the instantaneous rabbit-stage/perduring rabbit dispute, the case for division inscrutability is strong. I end by suggesting that such inscrutability, against initial appearances, should not be disquieting.

The fifth chapter turns to inscrutability arguments that are *prima facie* extremely disquieting: permutation-based and completeness-based arguments for *radical* inscrutability of reference.[7] I set out the former in an extremely general setting: a double indexed general semantics in the tradition of Montague and

[4] Alleged by Williamson (1994) among others.

[5] These are drawn from Evans (1975) and Fodor (1993).

[6] The generalization is due to unpublished work by Wolfgang Schwarz.

[7] I reserve the term for this case where the putative inscrutability of reference is 'total', and do not apply it to putative "Gavagai" style inscrutability.

The permutation argument can be found, in various forms, in Jeffrey (1964); Quine (1964); Field (1975); Wallace (1977); Putnam (1978a); Davidson (1979); Putnam (1981). The completeness/compactness arguments are mentioned by Putnam (1980).

Lewis.[8] I outline a small but significant extension of the inscrutability argument. The conclusion would be *indexical inscrutability*—it being indeterminate from one context to the next whether reference of a term remains the same. Finally, I outline two settings where the permutation argument breaks down: a structured-propositions framework and a Davidsonian truth-theory.[9] These do not, however, resolve the problem of radical inscrutability: but they relocate the issue in an interesting way.

Permutation arguments have one potentially serious defect: they characterize deviant interpretations in a way *parasitic* on an 'intended interpretation'. Arguments for radical inscrutability based on completeness results are slightly less general, but may ultimately prove more robust (as discussed in the final chapter of the thesis). I describe the way these work, and in Appendix B I give the details of completeness and compactness for a first-order language, and sketch the Henkin (1950) proof for completeness of an (extensional) type theory.

Part III: Against Radical Inscrutability

Davidson (1977) recommends we regard radical inscrutability as a discovery: there really is no fact of the matter about what our words refer to. Part three of the thesis considers reasons why this view might be unstable. I begin by outlining a number of putative reasons for rejecting radical inscrutability (first amongst them, the incredulous stare it generates). I then turn to extended discussion of two objections.

In chapter 6 "Lexical semantic beliefs" I put forward what I take to be a more serious concern: that radical inscrutability would be inconsistent with a 'cognitive' account of understanding. Drawing on work by Richard Heck, I outline a case for ascribing to speakers beliefs about the truth-conditions of sentences. This rests on what we need to do in order to represent speaker's linguistic actions as rational. Pointing to cases of subsentential speech acts, I argue that the same considerations implicate beliefs about reference in this picture of language use. I finish by outlining one response: accepting the need for beliefs about reference, but claiming that we can have (true) beliefs about the references of our words, due to a matching inscrutability in the contents of our thoughts, construed as expressions in 'mentalese'.[10]

Chapter 7 "Good inference and context" picks up indexical inscrutability result of chapter five, and argues that it will imperil the link between implication (what follows from what) and inference (what it is good to infer). I outline a puzzle for the Kaplanian treatment of consequence for indexical languages, and show how a certain modest proposal can resolve it. The modest proposal, however, involves the presupposition that the reference of an expressions is constant over time. I respond to objections to my account, and offer my own objections to alternative accounts.

Part IV: Avoiding Inscrutability

The arguments for radical inscrutability are based on the assumption that every semantic theory that 'fits' a certain range of data about the truth-values or truth-conditions of sentences is equally good, by interpretationist lights. In the final section of the thesis, I explore what happens when we add additional constraints to the selection of the meaning-fixing theory.

Chapter Eight "Eligibility" begins with methodological discussion of what sorts of supplementations of the interpretationist picture are philosophically defensible. The most promising *principled* response to inscrutability concerns I find in the literature is that offered by David Lewis (1983a, 1984): the 'Eligibility' response. I show that Lewis' 'eligibility constraint', and in particular, some *prima facie* peculiar aspects of it, is motivated directly from Lewis' interpretationism, when combined with several theses to

[8] In the main text, I develop the permutation argument for a 'pure categorial language'; in appendix C I extend this to a version of a λ-categorial language. (For the distinction between these frameworks, see Cresswell (1973)).

[9] For the former, see Soames (1989); for the latter, Davidson (1967) and Larson and Segal (1995).

[10] As noted below, it may be that the beliefs are only 'pseudo-semantic'—the relevant factor being that they can discharge the required theoretical role of genuine beliefs about reference.

which he is independently committed. This makes for an attractive package in the abstract, for it makes clear why the eligibility response is not ad hoc, but rather flows naturally from the interpretationist framework.

The proposal faces severe limitations. I argue that, in the case of global descriptivism, a 'revenge' problem arises. Inscrutability may be dealt with, but there are non-sceptical scenarios where the 'intended' interpretation is demonstrably *less eligible* than the kind of model we can build out of the natural numbers using Henkin-style techniques. I offer a way out to the Lewisian: if they except Armstrongian Universals at a macroscopic level, the revenge problem dissipates.[11] To avoid inscrutability in the Lewisian fashion, then, we may have to forswear microphysicalism.

[11] To equal effect, they could adopt what plays the role of 'emergent' sparse properties in alternative inegalitarian property ontologies such as resemblance nominalism, trope theory or naturalness-Primitivism.

Part I

The framework

Chapter 1

The semantic project

This chapter has two aims. In the first section, I introduce and motivate what I call ‘the initial semantic challenge’: to systematize the truth-conditions attaching to sentences. I give reasons for moving to a systematic theory that attributes semantic properties to individual words, rather than resting content with a mere listing of the truth-conditions.

I then turn to questions of the proper ambition of semantic project, and the nature of the resources we are to exploit when addressing it. I use Quine’s distinction between theories of reference and theories of meaning to focus this discussion. The conclusion is that, so far as the initial semantic challenge is concerned, both in ambition and execution we can remain within the theory of reference, so long as (contra Quine) modal resources form a legitimate part of this domain. This leaves open the question of the relation between a semantic theory and an account of the cognitive significance of language. I finish this section by outlining a number of ways in which the two projects may be related.

I turn next to the topic that forms the theme for this thesis: the *problem of intentionality*, conceived as the challenge to give a broadly reductive account of representation properties, including mental content and semantic content. *Metasemantic theories* focus on the latter element. I briefly describe how a familiar account—the causal theory of reference—fits into this setting, and introduce some strategic issues facing any metasemantic account.

Our focus in this thesis is on a particular metasemantic approach: *interpretationist* theories. I introduce the ‘two-step’ form common to such theories, and sketch three versions: David Lewis’ convention-based account; global descriptivism; and Donald Davidson’s radical interpretation. I finish by emphasizing the flexibility of interpretationism: in particular, there is no reason to suppose that it cannot be extended from natural languages to the language of thought; and no reason to think that it is committed to any objectionable semantic holism.

1.1 Semantic properties

The first half of this chapter discusses semantic properties. We start from the assumption that *sentences* have semantic properties: that they are systematically true and false in various circumstances. Do we have any reason to attribute other semantic properties?

We first look at reasons for positing, and developing a theory of, the semantic properties of individual words. What prompts the move from noting the distribution of truth-conditions to discerning, for example, *reference* relations between words and objects?

The second part of this chapter looks at the relationship between the project thus motivated, and the attempt to theorize about the fine-grained 'meaning' of words. I argue that the two projects are at least in principle separable; and outline some of the ways in which they may interact.

1.1.1 From truth-conditions to reference.

I take the following two principles as my starting point.

1. Given a sentence *S*, in one's own language, or in the language of another, one can evaluate whether what *S* says holds in various counterfactual situations. That is, sentences have *truth-conditions*.

2. At least in part, these truth-conditions are systematically sensitive to the (syntactic) parts of the sentence.

Though philosophers have questioned them, and there are certainly refinements and qualifications to be undertaken, to abandon them altogether would be a major revisionary project.[1] In what follows, I assume that (1) and (2) are correct.[2]

What do (1) and (2) describe? Effectively, they tell us of a range of facts that display a certain holistic interlocking character. Three instances of the kind of truth-conditional facts mentioned in (1) are the following:

- 'Billy runs' is true in all and only actual or counterfactual situations where Billy runs

- 'Susan runs' is true in all and only actual or counterfactual situations where Susan runs

- 'Billy walks' is true in all and only actual or counterfactual situations where Billy walks

Such facts are not unrelated. When we substitute one name or predicate for another, there is a predictable relationship between the truth-conditions of the initial and the resulting sentence. The situations that make-true 'Billy runs' and those that make-true 'Billy walks' both feature *Billy* doing something.

[1] One such revisionary project would be a pure 'disquotationalism', where a truth predicate may be introduced stipulatively through the laying down of the scheme:

'*S*' is true iff *S*

The resources thus introduced will be restricted in application to ascriptions of truth in one's own language. A pure disquotationism would insist that this was the *only* legitimate concept of truth. Consequently (1) would have to be given up (cf Gupta and Martínez-Fernández, 2005). Moderate disquotationists introduce a "use-dependent" notion of truth, exploiting translation into one's own language; they can therefore accept (1). See essays 3-5 of Field (2001b) and Field (2005); Loewer (2005); McGee (2005b). Such moderate disquotationalism is discussed below, as one way of avoiding the morals that I draw from (1) and (2).

[2] One important refinement is that we should distinguish what it is for a sentence *to be true when evaluated at* a situation *w* from its *being true as uttered in* that situation. This distinction is particularly important when analyzing indexical sentences (Kaplan, 1989b; Stalnaker, 1978). For example "I am here now", as uttered by me, is false when evaluated with respect to a situation in which I have moved from the spot in which I currently stand (that is what makes the sentence *contingent*). However, it is tempting to think that in all situations in which that sentence type is uttered, it will be true. Which of these two (1) and (2) are concerned with depends on how we are to understand the notion of 'what is said' by a sentence. See Stanley (2002).

The situations that make-true 'Billy runs' and those that make-true 'Susan runs' both feature someone *running*. Moreover, such connections support counterfactuals: Were it the case that a given sentence *S* had some alternative truth-conditions, we expect knock-on consequences for the truth-conditions of sentences other than *S*. For example, given the syntactic properties of language, had the string 'Billy runs' been true exactly in those situations where *Maxwell* runs, then 'Billy walks' would have been true exactly in those situations where Maxwell walks.

So far, I have simply described a range of systematic and counterfactually robust connections between the truth-conditions of sentences. It would be a *further* step to account for these connections by postulating an underlying level of semantic properties of individual words: for example, names *referring* to objects, and there being conditions under which a predicate *applies to* an object. Yet the move to this level is extremely natural. We want to say that 'Billy runs' and 'Billy walks' are each made-true by situations where Billy is doing something, because (A) 'Billy' refers to the individual Billy; and (B) generally, a sentence of the form '*N* walks'/'*N* runs' will be true in those situations where the individual referred to by *N* is walking/running respectively.

The challenge to give a theory of the semantic properties of words, from which we can derive the truth-conditional properties of sentences, I call the *initial semantic challenge*. If the challenge is accepted, nothing further needs to be said to motivate the project of *semantics*, understood simply as addressing the initial semantic challenge.

Since much of this thesis is taken up in examining the problems and puzzles that arise in giving a philosophical account of the facts invoked by this further level of theorizing, it is worth asking why we are moved to theorize at this level. Why should we take this 'further' step? Why not rest content with merely noting the systematic and counterfactually robust connections between the truth-conditions of sentences?[3] (For the time being, I assume that any 'further' theorizing here will take the form of ascribing referential semantic properties to words. There are those, such as Dummett (1991) and Field (2001a) who would propose a different kind of substantive theorizing, which I briefly discuss in a subsequent section).

The source of the challenge I: autonomous explanation

I have sympathy for the following description of the situation. Quite generally, we explain observed patterns by postulating an underlying system from which the pattern can be derived. The underlying system, if it were present, would allow a powerful *explanation* of the data. If asked why we postulate any such system, rather than presenting a 'theory' which simply lists the data, the reply is that on grounds of simplicity, unification, predictive power and so on, an underlying system provides *better explanation* than would a mere listing of the data. The data supports the postulation of an underlying system by *inference to the best explanation*. In the case at hand, the theory which ascribes referential properties to names and predicates—semantic theory—is explanatory of the pattern of truth-conditions of sentences in a way that a mere listing of the truth-conditions is not. Inference to the best explanation thus pressures us to take the 'further' step to systematic semantic theorizing.

The idea that semantics is in this way autonomous is well expressed by Scott Soames:

> ...one should look to [semantics] for an explication of the representational character of language. The central semantic fact about language is that it is used to represent the world.

[3] Foster (1976, p.2) argues that no substantive semantic project is motivated unless it has the ambition to account for speakers' competence in a language. Absent this, Foster suggests one might as well put forward a semantic theory that consists of the infinitary schema: $T(S) \leftrightarrow p$, where, in each case p is the object-language translation of the metalinguistic p. I have nothing against the project of accounting for linguistic competence: we discuss this below. The question here is whether Foster's claim that *only* linguistic competence could motivate substantive semantics. I argue that this is not the case.

> Sentences do this by systematically encoding information that characterizes the world as being one way or another. Semantics is the study of this information, and the principles by which it is encoded.
>
> (Soames, 1989, p.591)

I hold that (1) and (2), together with inference to the best explanation, motivate the initial semantic challenge and thus the project of semantic theorizing, in complete independence from any role in broader explanatory projects. To defend this view, and in particular the appeal to inference to the best explanation, would take us deep into issues of the nature of explanation, its methodological and epistemological significance, and applicability in relatively *a priori* disciplines such as semantics. I do not undertake this task here.[4]

The source of the challenge II: wider explanatory projects

Even if the 'minimal' and autonomous motivation for semantic theorizing just canvassed is rejected, there are various other pressures to indulge in semantic theorizing. The most famous of these is a putative connection between semantic theorizing and accounts of *linguistic competence*. Some indeed, see it as the *only* reason to engage in semantics (Foster, 1976). But even while resisting this extreme view, we may still agree that linguistic competence provides one reason to engage in semantics—much as an interest in fundamental physics may provide one reason for researching pure mathematics.[5]

Foster starts with the view that linguistic competence with *S* can be identified with knowledge of the truth-conditions of *S*. Since we are presumably competent with a potentially infinite variety of sentences, this immediately gives rise to puzzles about how we could have learned this infinite amount of information, and how we manage to store such information in our finite heads. An influential thought in response, abstractly stated, is that knowledge of potentially infinitely-many pieces of information is explicable, when we can identify a finite set of principles which *entail* the whole infinitary range of information. All we need do then is to credit agents with knowledge of this basic set (a finite learning/storage achievement) and the ability to derive consequences from this stored data.[6] One thing that semantic theory delivers is a way of encapsulating an infinitary range of data about the truth-conditions of sentences, within a finitely statable theory. Hence, one natural route to explaining how infinite linguistic competence is achieved by finite agents, is to credit the agents with knowledge of semantic theory.[7]

One need not buy into the 'cognitive' approach which identifies competence with knowledge of truth-conditions to see the attractions of building semantic theory into one's explanation of linguistic competence. Psychologically and neurophysiologically, as well as philosophically, it would be good to explain how an ability to understand potentially infinitely many novel sentences could be grounded in

[4] Notice, in particular, that accounts of inference to the best explanation that focus on *causal* explanation (cf. Lewis (1986b), Lipton (1991, ch.3)) will be problematic in the current, relatively *a priori* context. But there are other paradigms that look more promising: for example, the deductive-nomological model of explanation. See Lipton (1991, passim) for much relevant discussion and criticism.

[5] Soames (op cit) takes a view diametrically opposed to Foster. As quoted above, he holds that semantics is motivated autonomously as the study of the principles whereby linguistic representation is effected. But he also holds that semantic theorizing is ill-suited to the project of explaining competence. I discuss the relationship between the kind of semantic theories that Soames prefers and the project of accounting for 'cognitive significance' and linguistic competence later. See also Dowty (1979, ch.8.)

[6] See Chomsky on the analogous issue with syntactic knowledge.

[7] Foster and Davidson both state their aims a little more circumspectly: they want to explain linguistic competence by showing how it *could* be encapsulated by finite agents. They avoid saying that knowledge of semantic theory is what actually underlies linguistic competence.

For other philosophers endorsing the argument from a cognitive conception of competence (understanding as knowledge of meaning) to an interest in compositional semantic theories, see Dummett (1976), Evans (1981), Heck (forthcoming) .

finite minds. Hence the attraction of 'mentalist' assumptions about semantics in linguistics (cf. Davis and Gillon, 2004, ch.5); of the conception of semantic theory as describing the information held in our 'semantic module' (Larson and Segal, 1995), as well as of the claim within philosophy of language that understanding is knowledge of meaning. In all cases, interest in explaining linguistic competence provides a reason to be interested in semantic theory. Interest in any one of these projects would provide a motivation to take the 'further step' beyond noting relations between truth-conditions of whole sentences.

The source of the challenge III: counterfactually robust connections

The above discussion focuses on 'interlocking' truth-conditions of sentences. Our initial characterization of the data noted another feature: the interlocking of the truth-conditions of sentences is apparently counterfactually robust. Were 'Billy walks' to have truth-conditions other than those it actually has, the truth-conditions of other sentences would alter accordingly. A standard pattern for explaining counterfactual connections looks at the possible situations involving least disruption to the actual ways things are, but which render the antecedent of the counterfactual true; and then checks to see whether the consequent holds of the situation constructed.[8] We can fit the above counterfactuals into this pattern, by making appeal to the referential properties of 'Billy'. Given the way that reference and truth-conditions are compositionally related, letting 'Billy' refer to Maxwell, and holding all other lexical semantic facts fixed, we have a situation giving rise to the truth-conditions mentioned in the antecedent of the counterfactual. Moreover, this is plausibly the least disruptive way to accommodate the counterfactual supposition. But also, that situation makes true the consequent of the counterfactual. The counterfactual is thus straightforwardly vindicated.

If we were barred from appealing to the layer of lexical semantic facts, we would be left with a situation where, counterfactually supposing one sentence to have some property, we find knock-on effects for an (*ex hypothesi*) entirely unconnected property possessed by some other sentence. Brute counterfactual connections are rightly regarded with suspicion. Avoiding brute counterfactual connections, then, provides us with our final reason for moving to the level of subsentential semantic theorizing.[9]

To summarize: the systematic and counterfactually robust connections between the truth-conditional properties of sentences make pressing the initial semantic challenge in two ways. First, the systematic connections themselves demand an underlying explanation. I have a favoured view on what motivates the step from a mere listing of the truth-conditions of sentences, to subsentential semantic theorizing. However, I need not be over-committal on this point: a whole variety of explanatory projects, centrally those seeking to explain semantic competence, will require us to take the 'further' step. I need only maintain here that there is *some* well-motivated explanatory project of which subsentential semantic theory forms part.Second, the counterfactual robustness of these connections is itself barely intelligible if those connections are taken to be brute, rather than resting on an underlying basis of subsentential meaning.

[8] See Lewis (1979b). The challenge, of course, is to spell out what 'least disruption' amounts to. Lewis' account appeals to laws of nature. I would generalize this to an account that would allow appeal to laws of any special or fundamental science—including semantics.

[9] One response would be to attempt a similar explanation without appeal to *semantic* properties of individual words. For example, one might think there are causal-explanatory connections between words and parts of the world, and the truth-conditions had by sentences. The closest world counterfactual supposition holds is one where those relations differ; and this explains why the consequent holds.

The line of response is strategically good. However, the details look difficult. For example, the same kind of robust variation in truth-condition would appear to occur in discourse where we are not plausibly in any straightforward causal relationship with the items in question—for example, in discourse about abstracta. Thanks to Carrie Jenkins for pressing me on this point.

1.1.2 The theory of reference and the theory of meaning

We have just seen considerations pointing in favour of discerning semantic properties of words, as well as of complete sentences. We now turn to the question of the relation between the project thus motivated and another: that of accounting for the *cognitive significance* of words.

Quine's distinction

Quine distinguishes two kinds of projects within semantics, broadly construed:

> When the cleavage between meaning and reference is properly heeded, the problems of what is loosely called semantics become separated into two provinces so fundamentally distinct as not to deserve a joint appellation at all. They may be called the *theory of meaning* and the *theory of reference*. . . . The main concepts in the theory of meaning, apart from meaning itself, are *synonymy* (or sameness of meaning), *significance* (or possession of meaning) and *analyticity* (or truth by virtue of meaning). . . . The main concepts in the theory of reference are *naming, truth, denotation* (or truth-of) and *extension*. Another is the notion of *values* of variables.
>
> (Quine, 1953, p.130-1)

The distinction between these two projects can be brought out by considering the famous pair of sentences (Frege, 1892):

> Phosphorus is Phosphorus
>
> Hesperus is Phosphorus

The sentences have the same truth-value, and the words have the same reference. Indeed, orthodoxy has it that these two sentences agree in truth-conditions—since true identity statements are necessary, every situation in which one is true is a situation in which the other is true.[10] As far as the theory of reference is concerned, we can treat 'Hesperus' and 'Phosphorus' on a par, assigning them each the self-same semantic value.

On the other hand, it seems that we can know that the first sentence is true purely on the basis of our understanding of the terms involved; not so for the second. They are not 'cognitively' or 'analytically' equivalent. It is hard to see, therefore, how one could give a theory of understanding (semantic competence) that treated the two cases in the same way. *Prima facie*, the theory of meaning, responsive to such demands, will have to be more fine-grained than the theory of reference.

We want to know where to locate that project of semantic theorizing motivated by the initial semantic challenge. There are really two questions here. First, what should our ambition be? Should it be to underpin a theory of cognitive significance, or should its goals be characterized in terms drawn from the theory of reference? Second, what resources should we allow ourselves in formulating truth-conditions and prosecuting our explanatory goal: should we use the kind of rich vocabulary of the theory of meaning, or the more sparse equipment of the theory of reference?

The initial semantic challenge gives us our handle on this issue. It sets out the relevant kind of ambition: to systematize the truth-conditional properties of sentences. Our question then is what kind of properties truth-conditions are; and whether they are of the family that includes analyticity and synonymy

[10] See Kripke (1980) (originally published in 1972). For resistance to the orthodoxy, see Dummett (1973); Searle (1983); Lewis (1984); Jackson (1997).

(the theory of meaning) or whether they are more closely aligned to concepts of the theory of reference—reference, satisfaction and so forth.

We shall illustrate this issue by examining how it plays out with respect to two ways of formulating truth-conditions, corresponding to distinct frameworks for prosecuting the semantic project: *truth-theoretic semantics* and *intensional semantics*.

Davidsonian truth-theoretic semantics

Davidson's original project illustrates the way in which a theory can have the ambition to account for rich semantic properties such as cognitive significance, yet use resources drawn from the theory of reference. He recommends we formulate the truth-conditions of sentences in biconditionals such as:

> 'Schnee ist weiss' is true iff snow is white.[11]

Davidson then adapts Tarksi's theory of truth to show how axioms about the reference and application conditions of subsentential expressions allow one to construct a theory where such 'T-sentences' are derivable. Because *derivability* is at issue, there is potential a distinction in status between the following pair of truth-conditions:

(a) 'Hesperus is Phosphorus' is true iff Hesperus is Phosphorus

(b) 'Hesperus is Phosphorus' is true iff Phosphorus is Phosphorus

It would take an additional axiom of the theory (stating that Hesperus = Phosphorus) to let us interderive these two sentences. Hence, distinctions of cognitive significance might be captured by means of this syntactic way of presenting truth-conditions; and this without utilizing resources beyond those of the theory of reference.

Notice, however, that we are not *required* to exploit this fine-grainedness of the truth-theoretic setting. If our conception of truth-conditions draws no distinction between (a) and (b), then derivation of either would suffice to discharge the initial semantic challenge. Given such a coarser-grained conception of truth-conditions, the truth-theoretic framework still gives us a way of stating truth-conditions of sentences using resources purely from the theory of reference (and moreover, without invoking modal notions).

In sum, then, there are more and less ambitious ways of implementing the truth-theoretic account, according to whether or not we take truth-conditions to coincide with cognitive significance—this identification is allowed, but not mandated, by the Davidsonian framework. Whichever way we go as regards the *ambition* of the project, the *resources* exploited do not go beyond the theory of reference, strictly construed.

Intensional semantics and modality

The alternative way of formulating truth-conditions we discuss here is inherently modal. To begin with, notice that it is natural to take the *truth-conditions* of a sentence as captured by counterfactual conditionals (whether the sentence *would be true* if uttered in a given situation); or, more abstractly, as a matter of

[11] I use 'iff' throughout as shorthand for 'if and only if'. Throughout, I drop relativization to particular languages. The reader should have no problems adding such relativizations if required. (In fact, I do this for principled reasons: I thin that inter-linguistic homonyms should be treated as distinct *words* having relatively categorical semantic properties; rather than there being a single word which has different semantic properties relative to its use in two languages. For an account of word-individuation that cuts across orthographic or phonographic type, see Kaplan (1990).

the sentence's truth-values *with respect to actual and counterfactual situations*.[12] From this perspective, we naturally think of truth-conditions as 'world-conditions'—the set of conditions (i.e. possible worlds) where the sentence is true.[13]

A striking departure from the Davidsonian setting is the heavy use of modal machinery. Even formulating the ambition involves modal notions.[14] The natural way of implementing the semantic project, with truth-conditions formulated in this way, is through *possible worlds* or *intensional* semantics. Such a setting requires more than appeal to modal notions such as necessity, possibility, and (arguably) counterfactuals: it requires quantification over possible objects—talking donkeys, blue swans and merely possible situations.[15]

Two points are in order. First, that allowing quantification over possibilia allows one to define modal notions, and formulate possible-world semantics, in a purely first order way.[16] Second, that the increase in ontology must be defended; and if it is not construed literally, some reductive or deflationary story must be offered.[17] One concern is that this reductive or deflationary project will lead us back around to notions drawn from the theory of meaning. In fact, this is exactly how Quine conceived of notions of necessity—he held that a sentence 'necessarily p' should be understood as asserting the *analyticity* of the sentence 'p'.[18] It will be simplest if we finesse these issues, and assume *pro tem* an account of modality and possibilia that do not lead us back into the theory of meaning. I therefore take as a working hypothesis the *modalist* assumption that notions such as possibility and necessity can be taken as primitive, and need no independent explication; and combine this with either the kind of reduction of possibilia talk to talk of modality canvassed in Sider (2002); or the related ersatz-worlds proposal of Nolan (1997b). I would argue that these accounts need not appeal to *representational* resources.[19] If so, then intensional semantics involves no resources beyond the theory of reference plus modal primitives and abstracta.[20]

Given our assumptions, possible worlds semantics does not bring in the devices of theory of meaning. Just as with Davidson's project, one might *deploy* the semantic theory to account for cognitive significance on the basis of independent resources—in this case, the material of the theory of reference+primitive modality. As originally conceived, possible-worlds or intensional semantics did have the ambition to account for the cognitive significance of sentences, aiming to explicate Frege's notion of sense. The contention was that paradigmatic examples of cognitively significant identities such as 'Hesperus is Phosphorus' became false when evaluated at other possible worlds. These claims were famously rejected by Kripke, who put forward a persuasive case that such identities were *necessary truths*. Given

[12] The former understanding of 'truth-conditions' arguably would not capture the *semantic content* of the sentence; it would rather capture the 'diagonal content' or 'A-intension'. Either this notion or semantic content ('horizontal' or 'C-intension') are standardly represented in terms of functions from possible worlds to truth-values; but only the former is easily understood in couterfactual terms. For these distinctions, see Stalnaker (1978); Chalmers (1996); Jackson (1998).

[13] See Stalnaker (1984) and Lewis (1986c) for defence of the general idea of identifying truth-conditions with 'worlds-conditions'.

[14] One way of resiting this, though, would be to formulate truth-conditions counterfactually, rather than in terms of functions from worlds to truth-values, and then argue that counterfactual conditionals are not themselves modal. See Edgington (1995); Stalnaker (1984) for accounts of counterfactuals as projections of belief-revision policies.

[15] For recent discussion of the relation between modality and such possibilia, see Nolan (1997b); Sider (2002); Fine (2003).

[16] That is, the *semantics* for a multiply intentional type theory can be given in first order set theory with possible-worlds as urrelemente.

[17] Famously, David Lewis takes the existence of concrete cosmoi corresponding to all the ways the world could be, at least in part in order to give a reduction of modal ideology to the non-modal (cf. Lewis, 1986c, §1.2).

[18] This is Quine's 'second grade of modal involvement', alleged to involve use/mention confusion. Quine famously rejected even the notion of 'analyticity', so even the 'first grade' of modal involvement was too much for him.

[19] The reasons for this are the same as those presented in §2.3 below for similar 'fictionalist' proposals.

[20] Of course, it might turn out that best theory in the philosophy of modality recommends some alternative account—perhaps even one that recalls the pre-Quinean link between conventionality and modality (cf. Sider, 2003).

the Kripkean orthodoxy, the original ambition is not discharged.[21]

In the present context, the only impact of this observation is that the ambition of intensional semantics should not be to account for cognitive significance, in the first instance, but to focus on explaining truth-conditions. As it turns out, possible world semantics has proved hugely successful in giving satisfying theories of the truth-conditional structure of complex parts of language, for example modal discourse, counterfactuals and indicative conditionals, indexicals; as well as less obviously modal parts of language.[22] It is not obvious how to replicate these successes in less rich settings.

Summary

I have characterized the *initial semantic challenge*, and given several reasons for regarding it as pressing. I now distinguish two ambitions for a semantic theory: first, to account for the properties characterizable by the theory of reference; second, to account for the rich notions of the theory of meaning (centrally, the cognitive significance of sentences). Perhaps (as Davidson and the originators of intensional semantics thought) the project of giving the truth-conditions of sentences coincides with that of giving the cognitive significance of sentences. However, if this connection is broken, the motivation provided by the initial semantic challenge requires only that we pursue a systematic explanation of *truth-conditions*. The primary aim of the semantic project is not to explain cognitive significance: it would be a happy accident if truth-conditions and cognitive significance were to coincide.

The other question concerned the status of the resources used to formulate truth-conditions and prosecute our project. On at least one of the ways of formulating truth-conditions, they are formulated in modal or possibilist terms. Some (such as Quine) have claimed that modal notions should be analyzed via the theory of meaning. We have assumed *pro tem* that this is not the case, and under our supposition the initial semantic challenge and the framework for addressing it can be formulated in terms drawn from the theory of reference, supplemented with primitive modality.

Granting all this, we can still see the project of accounting for cognitive significance as interesting and important; and one that is intimately connected to theories of linguistic competence. We now turn to the relationship between semantic theory, motivated in the way described above, and 'meaning' richly construed.

1.1.3 The relation to meaning

Suppose that the oracle gave you the ultimate theory of the truth-conditions—every systematic variation is captured in an elegant way. Perhaps the truth-conditions the theory assigns to sentences match intuitive differences between the meanings of different words, as Davidson (1967) and others hope. But what if it doesn't? What if the truth-conditions assigned to "Hesperus is Phosphorus" and "Hesperus is Hesperus" are the same? Surely there's some interesting difference between the two cases. What can be done?

There are numerous approaches to this issue; some are listed below:

- *Reduce*. Look within the semantic theory that the oracle gave you, and see whether it contains the resources to account for the different standing of the two sentences. Even if what corresponds to truth-conditions within the theory does not track 'cognitive significance', some *other* semantic

[21] The recent trend towards two-dimensional semantics, in the hands of Jackson (1998) and Chalmers (1996), might be thought to resurrect a version of this original ambition. See below.

[22] Two famous works within this tradition familiar to philosophers are Lewis (1973b) on counterfactuals and Kaplan (1989b) on indexicals. Partee (1996) surveys the development and successes of Montague-style intensional semantics. Thomason (1974a) collects the writings of Montague himself on these topics. Lewis (1970a), Cresswell (1973) and Thomason (1974b) are clear introductions to the technical machinery. The introduction to Davis and Gillon (2004) surveys some different settings and extensions to the framework, and discuss rival approaches.

property of sentences might do so. First, if one is using an intensional semantics, semantic values of sentences are usually identified with something coarse-grained, such as a set of possible situations. But 'fine grained propositions' can be defined, by labelling a syntactic tree for the sentence with the semantic structure of the respective parts.[23] Second, are there semantic resources beyond *semantic content* (or '*C*-intensions'). Many theories include a notion of *character*—or systematic variation in content in different circumstances. Several would claim that there is a difference in the character (and the related "*A*-intension") of "Hesperus is Phosphorus" and "Hesperus is Hesperus". [24]

- *Bring in the cavalry*. If the semantic theory alone does not provide all the resources one needs to make discriminations between sentences with intuitively distinct meanings, then one can look to other independently motivated theories for help. Two examples are given below: they are each highly controversial, but give a flavour of the way that non-semantic theories might help to ease the theoretical burden on semantic theory proper.

  The first example is Robert Stalnaker's project, as set out in the essays collected in (Stalnaker, 1999a). There it is argued that a proper theory of communication can show how asserting sentences with the same semantic content can have interestingly different conversational effects—in particular, statements that are semantically trivial might nevertheless be informatively uttered. For Stalnaker, the idea that 'Hesperus' and 'Phosphorus' differ in semantic content is merely an illusion generated by the typically distinct pragmatic effects of sentences containing those names.[25] Here we have a recipe (albeit controversial) for relocating cognitive significance away from semantic theory and into a combined theory of semantics and pragmatics.

  Our second example is Fodor's appeal to non-intentionally individuated "modes of presentation". Fodor (1993) identifies these with words in a 'language of thought', which are individuated by computational role. The cognitive difference between 'Hesperus shines' and 'Phosphorus shines' is then located in the fact that the *vehicle* of the semantic content is distinct in the two cases. Again, given a distinctive and controversial background theory, differential cognitive significance no longer requires differences in semantic content.

- *Check your motivations*. What is the *need* for the differences in meaning? What theoretical function does making such distinctions serve? Is it possible to simply write off distinctions not capturable by one of the tactics above? Can we, that is, *eliminate* appeal to meanings from our best theory of the world, replacing any role they play by theoretically more tractable concepts?[26]

- *Supplement*. If one is dissatisfied by progress made through reduction and appeal to independent theories, and one holds that there are reasons for not eliminating the residual phenomena, then

[23] See Lewis (1970a). Compare the 'interpreted logical forms' of Larson and Ludlow (1993).

[24] *C*-intensions and *A*-intensions are two constructions within 'two-dimensional' intensional semantics. The two-dimensional framework is motivated to handle the behaviour of indexical expressions: see (Kaplan, 1989b; Perry, 1977). For the extension of this from paradigmatic indexicals to language in general see Lewis (1980); Jackson (1998); Chalmers (1996). For scepticism of this deployment of the two-dimensional framework, see Stalnaker (2001).

[25] For example, the differing informational content of speech acts differing only in substitution of co-referential names is discussed in Stalnaker (1978, 1984); the use of empty names is discussed in Stalnaker (1978); belief reports are discussed in Stalnaker (1987, 1988). The 'problem of deduction' is the main obstacle that Stalnaker sees (1984; 1991; 1999b). Lewis (1979a, 1970a, 1986c) outline further resources that, though rejected by Stalnaker himself, might be used within Stalnakerian project: *centred worlds* and *fine-grained propositions*.

[26] The famous eliminativist proposal is Quine's, who notoriously gives up the 'theory of meaning' as a bad job. Given his strict views on what counts as part of the theory of reference (the only remaining legitimate part of semantics)—and in particular, his association of modal notions with the supposedly discredited concepts of analyticity, Quine's eliminativism is far more radical than that canvassed here.

> one may need to *supplement* the theory with something else—a distinctive *meaning theory* to complement the *theory of truth-conditions*. This need not embed the theory of truth-conditions, but it had better be consistent with it.[27] The classic example of this move is, of course, Frege (1892), arguing for a theory of *Sinn* (sense) to supplement his theory of *Bedeutung* (reference). Field (1977) recommends that we supplement truth-theoretic semantics with an account of "conceptual role", which he goes on to characterize in a precise way.[28]

We have distinguished three motivations for undertaking the project of semantics: first, to give a unified account of the pattern of truth-conditions of sentences (the initial semantic challenge); second, to give an account of speaker's semantic *competence*; third, to give a description of the facts underpinning robust counterfactual truth-conditional connections between sentences.

I have urged that, at least if we do not share Quine's hostility to modal concepts, the initial semantic challenge can be formulated without appeal to terms drawn from the theory of meaning. Two approaches were then identified: Davidsonian truth-theoretic semantics; and model theoretic/possible-worlds semantics. The methodological stance of limiting one's ambitions to accounting for truth-conditions of sentences does not foreclose a more ambitious project of accounting for cognitive significance: we have just seen several ways in which the two may relate.

The ultimate semantic theory an oracle might give us would seemingly describe *semantic facts*—that words refer to particular bits of the world; that sentences represent that the world is thus-and-such a state. For the remainder of this chapter we will be looking at the philosophical puzzles that emerge in accounting for such properties and outlining the kind of foundational account that will be principal focus of this thesis.

[27] See the theories of meaning within linguistics described by Davis and Gillon (2004).

[28] The account of sameness of conceptual role, or *equipollence*, there characterized, is not easily extended beyond an analysis of *intralinguistic* synonymy. This contrasts sharply Frege, who placed great importance on the shareability of sense; and those (such as Stalnaker) whose basic interest is in the study of communication, rather than narrow psychological role. There is room in principle, therefore, for a three-way distinction: between truth-conditional semantics, meaning in the psychological sense; and meaning in the communicative sense. As I hope has become obvious, the relationship between these three projects is going to be a highly complex and controversial matter.

1.2 The problem of intentionality

Semantics itself is a research project that (in the first instance and with distinguished exceptions) is prosecuted by linguists rather than philosophers. Nevertheless, given a semantic theory of the Davidsonian or possible-worlds sort mentioned above, a distinctively philosophical question can be asked: what is the nature of the facts that it describes?

The general issue concerns the metaphysics of representational properties. Philosophers are interested in questions such as: in virtue of what does a sentence, picture, or mental state represent that the world is a certain way? This is the *problem of intentionality*.

One approach is to simply be brutally non-reductionist about the intentional, to maintain that there are emergent representational properties that cannot be explicated in other terms. Many are dissatisfied with such an approach. For example, Fodor writes:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in the face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else.
>
> (Fodor, 1987, p.97)

and Field:

> This doctrine, [which] might be called 'semanticalism', is the doctrine that there are irreducibly semantic facts. The semanticalist claims, in other words, that semantic phenomena (such as the fact that 'Schnee' refers to snow) must be accepted as primitive, in precisely the way that electromagnetic phenomena are accepted as primitive (by those who accept Maxwell's equations and reject the ether); and in precisely the way that biological and mental phenomena are accepted as primitive by vitalists and Cartesians. Semanticalism, like Cartesianism and vitalism, posits nonphysical primitives, and as a physicalist I believe that all three doctrines must be rejected.
>
> (Field, 1972, p.12)

One might reasonably wonder whether the dichotomy these authors present is genuine.[29] I want to outline an account that accepts the challenge as presented: the overall goal is a reductive account of the intentional. The specific project that concerns us here is a subproblem of the problem of intentionality focused on the semantic properties of language. What is it for a name such as 'London' to refer to something, or a predicate such as 'is large' to apply to some object? Such questions naturally arise if we think that there are semantic facts such as that the French word 'Londres' refers to London. *Prima facie*, if we are to take the kind of theories described above seriously, and regard them as true, it seems we are committed to take such facts seriously.

We can draw analogies to other areas of philosophy. If we take seriously the claims of ethics—that kicking puppies is wrong, or that *ceteris paribus* it is good if human welfare increases—then we need

[29] Fodor himself famously defends non-reductive physicalism about special sciences, so his appeal to the properties *physicists* list seems odd. McDowell (1978) discusses whether intentionality is a special case where a physicalistic worldview can be held compatibly with non-reductionism.

to explain what about the world, or about us, *makes it the case* that kicking puppies is wrong, or that human welfare is good. We need to explain the constitution of these ethical facts, and their relation to the physical. This leads to an interest in the field of meta-ethics, which addresses such questions. By analogy with the terminology in the ethical case, we can call the analogous foundational project, in the case of semantics, 'meta-semantics'.[30]

Of course, a metasemantical theory does not have to be a reductive one. Just as one might postulate irreducible 'non-natural' ethical properties, one might postulate irreducible representational properties. This is the position taken by Field's 'semanticalist'. Moreover, reductionism need not take the form of *reductive identification* of semantic facts with other (e.g. causally constituted) facts. Within the ambit of reductive accounts, I include fictionalist and anti-realist accounts that would give a substantive theory of how the apparently semantic emerges from the non-semantic. My focus will be on a certain class of reductive metasemantic theories ('interpretationist' theories). Before outlining the characteristics of this approach, I shall cover two more general issues: first, various ways of *avoiding* or *relocating* the challenge; second, the strategies available when constructing a reductional account of intentionality.

Who faces the challenge?

Metasemantics asks the question: What is the nature of the facts that semantic theories describe? What constitutes reference, satisfaction and the other relations that feature in classical semantic theorizing?

Who can avoid the specific foundational questions described above? Three kinds of theorist would appear to do so. Firstly, there are those, mentioned at the beginning of the thesis, who are *eliminativists* about the semantic—those who do not feel the pressure to engage in systematic explanatory theorizing about subsentential semantics, but rest content with a mere description of the differing truth-conditions of sentences; or more radically still, who do not think that there are systematically statable truth-conditions for sentences at all.[31] For reasons set out earlier, I reject this approach, and will not consider it further.

Second, there are those who, like the *deflationists*, deny that commitment to semantic talk about 'reference' and 'truth' commits one to semantic facts—who deny that there is a coherent project of *substantive* semantic theorizing.[32] This Quinean position has its own obligations and challenges. For example, as Field in his later work (1994; 2001a) develops it, a 'disquotational' (non-substantive) intralinguistic theory of truth and reference needs supplementation with a theory of *translation*, to account for how others who do not speak my idiolect can nevertheless be 'saying the same things' as me. We then need an account of what counts as a good translation, and the story here appeals to the kind of materials that might, in another setting, to give a metasemantic theory (Field says that S' will be a good translation of S if S and S' stand in approximately the same indication relations, and if the terms in S and S' have approximately the same inferential role.[33])

The other sort of theorist not to face the challenge in the form set out above is one who is *revisionist* about the form a semantic theory should take. For example, Michael Dummett, in (1991) and other writings, has argued that we should reject the 'classical' framework of reference and truth in favour of a constructivist alternative. Clearly, such a theorist is committed to the possibility of a constructivist

[30] I take this term from Stalnaker (1984, 1999a). There seems to be no universally accepted terminology in this area. 'Semantics' 'Pragmatics' and 'theory of meaning' have all been used to name the questions intended here; but such terms have all got other, more precise roles to play. I hope that the regimentation of terminology I use is not artificial.

[31] This is *not* the Quinean eliminativism about analyticity and meaning more generally. For, as discussed above, the more hygienic theoretical enterprise of the *theory of reference* (generously construed to include modality) allows the foundational questions can be posed.

[32] Typically, they think that the needed notions can be characterized intra-linguistically through the so-called T- and R-schemata. See, for example, Horwich (1990).

[33] See Field (2005); Loewer (2005) for a synopsis of the view. The original Quinean version had translation constrained by similarity of patterns of assent and dissent. See Quine (1960, ch.2.).

account of comparable detail and success as the more familiar classical alternative. Once such a theory has been developed, foundational questions can be put which bear comparison to those detailed above.

The two approaches just mentioned do not eliminate the metasemantic questions, but simply transform them. What makes one *translation* better than another? What story can one give about the foundations of new non-classical semantic theories? Though they avoid "straight" metasemantic questions, they face "relocated" ones. That is not to say that no progress can be made by such projects: in principle, the relocated questions may be more tractable than the original.

I mention such views only to set them aside. The challenge to be considered in this thesis accepts "straight" semantic theory, construed as describing robust word-world representational relations.

Head-first and word-first strategies.

In the work cited, Field and Fodor endorse a *causal* account of the reference of terms, at least for basic parts of language.[34]

One interesting thing about these authors is that they adopt a *word-first* strategy in their overall reductive account of intentionality. The foundational theory they demand for language will not make appeal to any intentional mental states for example. They can then supplement this story by a reduction of mental content to linguistic content, by supposing that the vehicle of mental content is itself a language—the language of thought. (See Field (1978), Fodor (1987, passim).) In this way a reductive account of linguistic content may be turned into an overall reductive account of the intentional.

The strategy contrasts with that adopted by other causal theorists. Kripke (1980) notes explicitly that his 'causal theory of reference' makes appeal to certain *intentions*. Such a causal theory, if put forward in the service of an overall reductive account of the intentional, *presupposes* some independent foundational story of mental content. Stalnaker (1984, 1997), developing the Kripkean causal theory, explicitly puts forward a *head-first* account: giving a reductive account of mental content, and then appealing to mental content when giving foundations for the content of language.

Alongside word-first and head-first strategies of Field and Stalnaker respectively, there are also 'no-priority' strategies, which hold that neither mental nor linguistic content can be fixed independently of the other. This was Davidson's preferred approach (Davidson, 1974).

A second strategic choice faces the metasemantic theorist, within their account of linguistic content. Do they, with Field (1972) and the other causal theorists mentioned earlier, see the semantic properties of sentences as built out of, and reducible to, the semantic properties of individual words (What Davidson (1977) calls the 'building-block' view). Or is some account in view that assigns content to sentences and then *derivatively* characterizes the content of words, as urged by Davidson (1973); McDowell (1978)? Or should we favour a *holistic* treatment whereby the semantic properties of individual words and whole sentences are fixed simultaneously?[35]

This disagreement over broad reductive strategy (head-first vs. word-first) cuts across disagreements over the 'direction of explanation' within the account of linguistic content. For example, Lewis (1994b), like Stalnaker, favours a head-first strategy; but his account of linguistic content (described below) is not of the building-block kind favoured by Stalnaker.[36]

The choice of reductive strategy is of enormous importance for the overall project of addressing the problem of the intentionality. Word-first theories are hugely ambitious, taking on the problem directly; whereas head-first theories allow only a partial reduction of linguistic content, and some independent

[34] See Field (2005) for discussion of his views on explicating 'primitive denotation', including non-causal elements. Fodor (1993) discusses how to give a foundational story beyond the paradigmatic cases where a causal theory seems appropriate.

[35] As will be seen below, I believe the latter to be the most promising way of developing the 'top down' Davidsonian picture.

[36] Their account of mental content is reasonably similar, giving a central role to decision theoretic machinery. However, Stalnaker (1984) appeals to causal connections to get an independent fix on the content of belief, whereas Lewis adopts a more holistic strategy.

story about mental content would be offered if the intentional is to be reduced to the non-intentional facts. When we discuss specific metasemantic theories below, the kind of strategy that they allow will be of special interest.

Those who take semantic facts seriously face *metasemantic* questions, about the nature of such facts. This is a subquestion of the general problem of intentionality—the metaphysical standing of representational properties. A dualism, that simply takes such facts as primitive, is a *prima facie* coherent view, but not a very attractive one. One can relocate the problem in various ways, but avoiding it altogether seems unlikely.

My project in the thesis is to examine a "straight" broadly reductive account, describing what underlies best semantic theory. We have outlined some of the strategic choices that face the theorist over how the reductive account of semantic facts fits into an overall reductive account of the intentional.

In the remainder of this chapter, we describe *interpretationist* metasemantic theories. These will come in both head-first and word-first varieties; but quite generally reject building-block approaches to reducing linguistic content. The alternative picture is escribed in the following section, and examined in detail in Chapter 2 .[37]

[37] I will not engage in criticism of the building-block approaches here. But I cannot resist expressing puzzlement over one feature. Let us suppose that the semantic content of all lexical items can be fixed in some fashion. In order to extract an account of the semantic content of sentences, one needs to appeal to compositional axioms. The semantic significance of composition *prima facie* cannot be taken for granted within a foundational theory. In general, semantic projection rules vary from one semantic setting to another. It is hard to see how we could be entitled to ignore such features within a metasemantic theory, and obscure to see how a building-block theory could handle them.

1.3 Interpretationism

Interpretationism is an approach to the metasemantic challenge under which a variety of accounts fall. As I shall develop the approach, it engages squarely with the Fodor-Field reductive challenge: to explain how representational properties can exist in a world that can be characterized exhaustively in non-representational terms.[38] However, it does not attempt to do this in the 'building-block' style characterized above. Rather, one first characterizes *holistically* what it is for a semantic theory to be in force, and then ascribes reference to terms and truth-conditions to sentences on the basis of what the semantic theory entails. In the next chapter, we shall discuss what metaphysical view of semantic properties such a 'theory shadowing' account suggests.

Once we have chosen a semantic theory, then we can say what it is for *e* to have semantic property *P*, in terms of what follows from the 'selected' semantic theory. The fundamental question facing such an account, therefore, is how this meaning-fixing theory is selected. It is here that we find the substance of interpretationist proposals. They characteristically adopt a certain two-step strategy: first, identifying a range of data correlating sentences with states of the world; second, saying that for a semantic theory to be in force in the population (to be selected) is for it to be the *best* theoretical account of the data.

Let us illustrate this with a toy case. Suppose that Billy and Jane speak a simple language. Intuitively, they have names for one another "B" and "J", and predicates for "wearing red", "wearing yellow" and "wearing green": "R", "Y" and "G". For the sake of our toy example, suppose they are invariably successful detectors of each other's clothing, and always speak their minds.

Now, causal theorists would look for causal links between the individual lexical items—say "B" and "J"—with items in the world. By contrast, the interpretationist first gathers the data:

"J R" is uttered when and only when Jane wears red

"J G" is uttered when and only when Jane wears green

"J Y" is uttered when and only when Jane wears yellow

"B R" is uttered when and only when Billy wears red

"B G" is uttered when and only when Billy wears green

"B Y" is uttered when and only when Billy wears yellow

Now given a toy semantic theory correlating "J" with Jane, "R" with *wearing red* and so forth, we can derive statements such as:

"J R" is true iff Jane wears red

and indeed, appealing to such axioms, we can derive a theorem corresponding to each data point.

If we suppose that Jane and Billy have competence with the semantic theory, and that they try to utter something just when it is true, then the little semantic theory we have constructed will *account* for this data. Our hypothesis, then, is that the designated theory will be the one that best discharges this role, accounting for the data about sentence utterances. A natural language contains far more complexity, and elements of the scenario described are implausible (they utter the sentences *whenever* the condition is met?). However, the principles will be the same. Write down correlations (identified in non-semantic terms) and then select that theory that (modulo pragmatic hypotheses) *best explains* this data.

[38] I rule that the 'dual aspect' position suggested by McDowell (1978) are incompatible with this ambition.

(To avoid possible confusion. I am not here thinking of an epistemological project, where certain theorists *have some data* and we ask what theory they would ideally choose. Rather the 'data' are simply all the facts of some characteristic sort, which no theorist need have access to. Compare the 'Humean' accounts of laws of nature, and in particular Lewis (1994a), where a similar non-epistemological relationship between 'data' and 'best system' is appealed to.[39])

Within this broad framework, there are many loci for variation. How shall we characterize the data? What form should it be presented in? What kind of semantic theory should we look for? What kind of pragmatic considerations are appropriate to appeal to? Are we allowed to appeal to intentional properties in the process (appealing to a 'head-first' approach to the problem of intentionality) or are more severe constraints in place? It is in answering these questions that the various forms of interpretationism get their distinctive character.

I shall briefly sketch three ways in which interpretationism can be developed. Each no doubt deserves an extensive discussion. We are concerned, however, with issues common to all such approaches. In chapter 2, we shall be looking at what metaphysical picture of facts about content they offer; and the much of the rest of the thesis is concerned with examining "inscrutability puzzles" that *prima facie*, arise for the same reasons in each one of these approaches. It is therefore appropriate to introduce the theories here succinctly, and to note differences as they become relevant in the course of our later discussions.

Lewis' convention-based interpretationism

Lewis' head-first reductive account of intentionality incorporates an interpretationism about semantic properties.[40] The first step is to identify *linguistic conventions* that govern utterances of sentences: these will provide the pairings of sentences with states of the world. A convention in Lewis' sense is a certain kind of regularity in action; one characterized by mutual expectations and preferences sustaining the regularity. Paradigm examples include the convention to drive on the left hand side of the road when in the UK, and the convention to meet with friends for lunch in a particular venue. There is nothing intrinsically superior about driving on the left rather than the right; there is nothing, we can suppose, favouring one venue over another; but the participants have an interest in co-ordinating their actions: they prefer to drive on the left *given that everyone else does*, or meet in the venue *if that's where other people will be*. In the jargon of game theory, they face a *co-ordination problem*. Lewis (1969) is a book-length treatment of the general notion of a convention based on this idea.

Informally, we have the idea that R is a convention iff the following three conditions are met:

1. REGULARITY: Everyone conforms to R
2. EXPECTED CONFORMITY: Everyone believes that everyone conforms to R
3. SUSTAINMENT: R is the solution of a (repeated) co-ordination problem.

The last condition can be broken down into (at least) three components:

(3) (a) COOPERATION: coincidence of interest dominates;

(b) EQUILIBRIUM: all would prefer that they themselves conform, on the supposition that the others conform;

[39] The relationship between interpretationism and Humean accounts of laws is discussed in Chapter 8.

[40] For the overall project see Lewis (1974a, 1994b). For the specifically interpretationist component see Lewis (1969, 1975, 1992). For the head-first account of mental content see Lewis (1974a, 1994b), Lewis (1986c, §1.4).

Lewis frames his views with a possible-worlds semantics, which dovetails neatly with his favoured way of describing mental content; but it seems that his account is largely independent of this assumption.

(c) ALTERNATIVE: there is some possible regularity R' uniformly incompatible with R, which, were it to obtain, would meet the above conditions. (so we have here a co-ordination *problem*).

Lewis then refines and liberalizes the basic idea expressed above to arrive at his final analysis. But the essential features are already present: conventions, not as explicit agreements, but as regularities sustained by characteristic attitudes within a population.

Linguistic conventions are then an application of this analysis. The basic resources are conventions of truthfulness, cashed-out in the Lewisian way:

> The regularity of *uttering S only if one believes p* in the behaviour of members of the population P in a serious communication setting[41] is a *convention of truthfulness* iff it is true that, and it is common knowledge in P that, in any serious communication setting:
>
> 1. Everyone utters S only if they believe p;
> 2. Everyone expects everyone else to utter S only if they believe p;
> 3. Everyone has approximately the same preferences regarding all (relevant) possible combinations of actions;
> 4. Everyone prefers that everyone utter S only if they believe p, on condition that at least all but one utters S only if p;
> 5. Everyone would prefer that everyone utter S only if q, on condition that at least all but one utter S only if q,
>
> where q is some proposition distinct from p.[42]

The pairing of sentences with states of the world that interpretationism requires is thus provided for: S is paired with the proposition p iff there is a convention to *only utter S if one believes that p*. Clearly, this requires a 'head-first' approach to the problem of intentionality, given that it involves appeal to beliefs and desires. Indeed, Lewis explicitly adopts such an approach, in a form relatively close to Stalnaker's.

Lewis (1975) exploits linguistic conventions to develop an interpretationist metasemantic account. Whether or not a semantic theory (what he calls a 'grammar') is "used by" a given population, turns on which assignment of propositions to sentences (what he calls a 'language') is "used by" that population. This in turn is fixed by the linguistic conventions that prevail:

> I would say that a grammar Γ is used by P if and only if Γ is a best grammar for a language $\mathcal{L}$ that is used by P in virtue of a convention in P of truthfulness and trust in $\mathcal{L}$; and I would define the meaning in P of a constituent or phrase... accordingly.

[41] This is one of the alternatives Lewis (1975, p.183) suggests in response to issues about non-literal use of language. A *serious communication setting* with respect to a sentence S and proposition p obtains whenever:
it is true, and common knowledge between a speaker and hearer that

1. The speaker does, and the hearer does not, know whether p;
2. the hearer wants to know;
3. neither the speaker nor the hearer has other (comparably strong) desires as to whether or not the speaker utters S

[42] This formulation is based on those found in (Lewis, 1969). In (1975), Lewis complicates the account by adding 'conventions of trust': the conventional regularity of *forming the belief that p in response to hearing someone utter S*.

Further, more complex, approaches to specifying appropriate linguistic conventions are possible. Griceans, for example, might wish to adopt interpretationism about *semantic meaning* by means of conventions of individuals to speaker-mean p when uttering S (cf. Schiffer, 1972). Avramides (1997) endorses this kind of proposal. Not only does it extract a notion of the 'timeless meaning' of sentences from the representational properties of sentence-tokens the Gricean has available, but it will underpin ascriptions of subsentential meaning, which otherwise have no obvious place in the Gricean framework.

(Lewis, 1975, p.177)[43]

Here Lewis' notion of a semantic theory 'used by' a population corresponds to my notion of a 'selected' semantic theory.

In Lewis' account of what makes a 'language' correct, characterized in terms of conventions, we find the characteristic first step of an interpretationist metasemantics—identification of sentential data. In his account of the correctness of a 'grammar' (semantic theory) we find the second component—lexical meaning-facts fixed by the best theory of this data.

Global Descriptivism

Lewis (1984), drawing on Putnam (1980, 1981), describes an alternative to his convention-based approach—*global descriptivism*. First, one must construct, for language as a whole, a 'term-introducing theory'. Lewis says that this will be "total theory": the set of all the sentences that normal (or perhaps ideal) agents would endorse.

Lewis is extremely unspecific about exactly how "total theory" is to be identified; but then, he is perhaps not himself endorsing the theory, but rather offering it as a reconstruction of the view described as 'standard' by Putnam (1980). Global descriptivism therefore delimits a *class* of views. One in particular is often associated with Lewis. This sees it as a generalization of the so-called 'Ramsey-Carnap-Lewis' treatment of theoretical terms (Lewis, 1970b). Here, the first step is to identify a 'folk theory' of some phenomenon (say, heat, or mental phenomena, or inheritance). These are supposed to be platitudes with which no-one should disagree. In the case of mind, this might include principles such as: "If someone hits you hard, and you are paying attention, you will feel pain"; and "If you are feeling pain, then unless distracted you will tend to wince and groan". The introduction of theoretical terms is supposed to be accomplished by collecting such platitudes, formulating them efficiently and then transforming the resulting *folk theory* into a definition, say, of mental vocabulary, by the technique of 'Ramsification'.[44]

Suppose that we collect *all* the platitudes from *every* walk of life, and formulate them into a global 'folk theory'. Then we have exactly the sort of "total theory" that we need to introduce global descriptivism. On this implementation, the interpretationist's data is the *uninterpreted* global folk theory.

Where does this version of global descriptivism stand strategically? Lewis does not say; if platitudes can be identified in a non-intentional way, it can be a word-first view. Easier, of course, would be to identify platitudes given prior information about which sentences are *believed* to be trivial; but this surely would only make sense within a head-first view.[45]

Next:

> The intended interpretation will be the one, if such there be, that makes the term-introducing theory come out true.

(Lewis, 1984, p.60)[46]

Again we find a two-step strategy. The first interpretationist step is to identify an appropriate set of sentences—the global 'total theory' to which Lewis refers—perhaps global folk theory. The second interpretationist step is to find a semantic theory that renders all (or enough) of these sentences true.

[43] Page references are to the version collected in Lewis (1983b).

[44] For an account of this, see Lewis (op cit).

[45] Lewis, as already mentioned, favours a head-first approach; but in presenting global descriptivism he may well wish to describe an approach that is non-committal on this point. Indeed, he indicates at points in Lewis (1983a) that he regards it as a word-first approach, and objectionable on that count.

[46] Page references are to the version collected in Lewis (1999).

The semantic framework that Lewis has in mind seems to be model-theoretic in character. Beyond this, presumably the only constraints depend on what resources are required to get a sensible theory making total theory come out true; with purely first-order vocabulary, we might get away with a first order language; whereas with modal vocabulary involved, a Kripke-semantics may be required.[47]

Davidsonian radical interpretation

In a series of papers in the 1970's, Davidson developed the kind of interpretationist approach we are examining.[48] The interpretation of Davidson's views that I favour is far from uncontroversial, for I will look on him as putting forward a purely *metaphysical* or *metasemantic* project; whereas it is often thought that the key element of Davidson's radical interpretation is an epistemological story about how one interprets others in everyday situations.[49] Having noted these reservations, I will for convenience set exegetical issues aside and speak as if Davidson's view were straightforwardly metasemantic in ambition.

Davidson takes as his starting point descriptions of situations described in non-intentional terms. Individuals behave in various ways, utter strings of sounds in certain circumstances, and the rest. Davidson supposes that, on this basis, we can identify which sentences a subject "holds-true"; or which she "prefers-true" to others.[50] Davidson is more radical than Lewis, in that his ambition to reduce both facts about the content of attitudinal states, and semantic facts, to the spartan basis of the pattern of sentences held- or preferred-true. His strategy is neither head-first nor word-first: he aims to get both linguistic and mental content simultaneously.

In its original version, radical interpretation proceeds by finding generalizations about the circumstances in which sentences come to be held-true. For our toy language above, for example, the generalization might be that one holds-true "J G" whenever Jane is wearing green in the vicinity. Taking this as one's starting point, one can start to compile a list of initial hypotheses linking sentences with truth-conditions.

> "J G" is true if and only if Jane is wearing green

The ambition is to get a workable set of 'T-sentences' that will *transform* information about sentences held-true, to information about the beliefs an agent holds; and which *in addition* form the target theorems for a semantic theory to generate.

There is a complex story about how the initial set of generalizations is to be revised and improved until an acceptable final set is reached. The process is regulated by so-called principles of 'charity' or 'humanity', which admonish the interpreter to maximise the rationality and truth found in the subject's behaviour and beliefs; the eventual semantic theory and assignment of beliefs is one that *optimises* such virtues.[51]

[47] Contrast with the convention based approach above, where the pairing of sentences with propositions (sets of possible worlds) forces one into a possible-worlds semantics setting from the beginning.

[48] See the papers collected in Davidson (1984). Davidson (1980) marks an important development in Davidson's account.

[49] Heck (2005b) defends the kind of view I take of the ambition of Davidson's project. Lepore and Ludwig (2005) examine in great detail Davidson's writings, favouring a more epistemological reading. I find Davidson's own writings obscure on this point.

[50] This is often held to be an encroachment of the intentional into the basis of Davidson's radical interpretation. See Heal (1997). I am less sure of this. It seems to me that it is plausible that one can identify attitude *kinds* independently of the content of attitudes. See Fodor (1987, ch.2.) for the importance of this distinction.

Believing-true is plausibly a *kind* of attitude (a sub-kind of belief), so it may be that we can use the above to get a grip on which sentences are held-true. Fodor's setup allows a straightforward characterization those sentences in the language of thought which are 'held true' (or, as he might put it, those sentences that 'are in the belief box'). Since Davidson is interested in natural language, he will have to appeal to the manifestation of attitudes in utterance, and it is not clear how the story will go. In any case, it seems clear to me that the challenge is qualitatively different from that of identifying the content of ordinary beliefs.

[51] Lewis (1974a) and Lepore and Ludwig (2005, pt II) contain extensive discussion of the details of Davidson's project.

Davidson (1980) extends the procedure in an innovative way. By focusing on data about preferences-true, and exploiting the decision theoretic determination of cardinal attitudes from ordinal attitudes, he is able: (1) to describe how (in an idealized setting) to identify truth-functional connectives within a language; (2) to describe how to identify *degrees* of holding-true and preferring-true, using the methods of Jeffrey (1965). This richer basis is then used to give a fuller account of how radical interpretation of a *full language* (beyond the core of 'occasion' sentences) might proceed.[52] (The new ideas introduced in this article can be elaborated in non-Davidsonian ways, too.[53])

However the details work out, we have the characteristic patter: primitive semantic data is given (the finalized set of T-sentences); and the correct semantic theory is picked out as one that generates the data. The choice of T-sentences as data is, of course, not accidental: for this is the characteristic output of Davidson's favoured form for a semantic theory—truth theoretic semantics.[54]

Flexibility

Particular implementations of interpretationism are committal. For example: Lewis' convention-based approach is appropriate to *natural languages* rather than *languages of thought*; and to *communal* language, rather than individual idiolects. Davidson similarly focuses on natural languages; though his approach is more sympathetic to idiolects than is Lewis'. Different authors construct their theories with different semantic frameworks in mind.

Some of the characteristics mentioned above are inessential to the overall strategy; and of course interpretationism as an *overall* strategy is compatible with a wide range of starting points, simply because of the varying nature of the views.

There are two points, however, that might be thought to be theoretical commitments of any interpretational theory. These are a focus on *natural* languages, rather than languages of thought; and a commitment to a form of 'meaning holism' that some find objectionable.

It is, of course, extremely controversial whether there is a 'language of thought'—a syntactical structure with recognizable names, predicates and logical connectives, which underlies mental computation. It would be a bad thing for an approach to reducing the intentional if it were *incompatible* with this hypothesis, or could offer no story to support it.[55]

However, there seems no real obstacle to constructing an interpretationism that works in that setting. Consider, for example, the 'head-first' strategy endorsed by Stalnaker, Lewis and others. The head-first approach uses causal indication relations, decision theoretic constraints, and similar factors, to determine the *overall* content of an agent's belief and desire states. Surely, there must be some neurological vehicle for this content—some information-storage mechanism. Though Stalnaker and Lewis proclaim their *neutrality* on this score, they do not *reject* the hypothesis that the information is stored in sentence-like structures.[56]

On this view, sentences in the language of thought within a subject's 'belief box' would collectively represent the world being a certain way. The challenge for an interpretationist is then to assign content

[52] See Lepore and Ludwig (2005, ch.16) for discussion.

[53] One simple non-Davidsonian proposal would be to use the detailed data about sentential attitudes to evaluate the success of pairings of sentences with propositions: one would attempt to optimize the rationality of the agent's non-linguistic actions, along the lines suggested by Lewis (1994b). The optimal pairing would then serve to identify the agent's attitudes, and further to provide the target data for a semantic theory.

[54] For a contemporary introduction from a linguistics perspective, see Larson and Segal (1995). Lepore and Ludwig (2005) discuss Davidson's version.

[55] Of course, one might agree with a language of thought as a hypothesis about cognitive functioning, but not think of the 'language' as having a compositional semantics—this would be directly compatible with the Stalnaker-Lewis holistic mental content. This would not satisfy Fodor (1993), however.

[56] Lewis (1994b). See the general discussion in Braddon-Mitchell and Jackson (1996).

to the individual sentences, and to the *lexical elements* of those sentences. One option is to appeal to a variant of global descriptivism to accomplish this: one would identify 'total theory' with the sentences that the subject is disposed to put in the belief box. (Alternatives are possible: it seems to me, for example, that ideas from Davidson (1980) do extremely well when adapted to this setting. The key observation here is that Davidson's basic notion of a sentence being *held true* can be cashed-out, within a Fodorian setting, in terms of *S* being tokened in 'the belief box'.) On the other hand, it is hard to see how to apply the Lewisian convention-based approach to this setting. There is little sense in which we have preferences about one another's tokening of mentalese sentences, as we do about each others utterances in a communal language. I hope that even this quick sketch will show that there is nothing in interpretationism incompatible with a metasemantics for the language of thought.

To what extent is the interpretationist committed to holism? Well, it is undoubtedly the case that *metaphysically*, it pictures semantic facts as determined by the best theory of *all* language use. Moreover, since content is assigned to expressions via a semantic theory for a language as a whole, semantic properties such as 'having reference' will be 'anatomic' in the sense of Fodor and Lepore (1992): necessarily, if one word refers, many others do too. However, it is the kind of holism or anatomism that these authors call 'weak' rather than 'strong'—for a word to have reference, we need only that *there be* other words with references. Strong holism would require that there be words meaning such-and-such (in English, say), such that if 'London' is mean the same in *L* as it does in English, the same must go for those words. But, on the contrary, on the interpretationist proposal there is every reason to think that words in languages with only small areas of overlap can share content. For, according to the interpretationist, the content that a term has is that assigned to it by the relevant theory. If the best semantics for British-English declares that 'London' refers to London; and the best semantics for Martian declares that 'ZXQU' refers to London; then the two words have the same content.[57]

[57] Given that Fodor and Lepore (1992) are concerned to argue that holism does *not* follow from a variety of theories (including Davidson's and Lewis', which we have briefly mentioned above) there need to be no dispute between us. It is significant that a clearly 'holist' approach to meaning-fixation does not motivate the strong holist theses that interest Fodor and Lepore. Indeed, their 'strong holism' seems so strong it puts into doubt their claim that "almost everyone is a holist" (p.32). Certainly Lewis, as I read him, is committed to no such claim.

In fact, some local versions of 'strong anatomism' seem uncontroversial. If 'Brutus killed Caesar' means what it does, then, one might reasonably think, 'Caesar killed Brutus' must also mean what it does. This certainly follows from the kind of proposal which I am interested in, but it is far from the meaning holism (or indeed 'molecularism') in the sense that Fodor and Lepore wish to engage.

Chapter 2

Reductive paraphrase

We are challenged to give a foundational story about semantic notions such as reference. As prefigured in the previous chapter, the metasemantic project that we are interested in is an interpretationist one: and we have seen several shapes that such a project might take. What *kind* of proposal is this? I shall outline the enterprise as one of providing a *reductive paraphrase* for discourse about the semantic. This paraphrase will pair sentences containing semantic terminology such as 'says' 'refers' 'has truth-conditions', with sentences that are free of such terminology. Comparable projects focus on paraphrases that reduce *ontology*: the elimination of appeal to possible worlds, abstracta, composite objects etc. We are here interested in reducing a certain kind of *ideology*: i.e. showing that certain distinctive vocabulary doesn't have to be taken as primitive. The basic goal of the current chapter is to delimit the form and detail of the reductive paraphrase associated with interpretationist metasemantic theories. I will also show how the proposal can escape various charges of circularity.

Confusion may reign if we do not first get clear on the nature of the project being undertaken. What are the aims of reductive paraphrase? What are the constraints on its success? Do we need to offer some kind of analysis of what the discourse in question *means* (a kind of conceptual or semantic *analysis* of the discourse in question)? If not, what are we up to?

2.1 The framework: theory-shadowing paraphrase

A common tactic when trying to avoid metaphysical commitments is to offer a paraphrase of apparently committal discourse. The paraphrase will pair sentences with non-committal surrogates. Of course, the surrogate sentences may have commitments of their own: reductive paraphrase may take the form of a trade-off. Thus Lewis (1968) offers a paraphrase of modal claims into a combination of possibilist quantification, and the invocation of certain similarity relations among possibilia.

There is a certain picture of the goal of reductive paraphrase that I want to resist. This has paraphrase 'uncovering' *what was really meant all along*. Manifestations of this picture include the thought that sentences paired by a reductive paraphrase should have *the same meaning*, or that paraphrase should take the form of semantic analysis of the discourse in question.[1] I think this picture is problematic in a number of ways. It lies at the heart of famous 'symmetry' objections to the paraphrase project: since synonymy is symmetrical, why should we take a paraphrase to deflate the ontological pretensions of the original (apparently committal) discourse, rather than uncovering previously hidden ontological costs of the surrogate (apparently non-committal) discourse? Moreover, the picture seems to burden us with a range of objections from sources that should be irrelevant. For example, suppose that one were able to give a paraphrase of arithmetical talk into terms that (a) had no apparent commitment to abstracta (b) were such that the surrogate could discharge the theoretical role of arithmetic, e.g. within science. The result would seem extremely philosophically important: exactly the sort of thing to make nominalists rejoice. However, it may not be plausible as semantic analysis, for sociological reasons. We can imagine a community who insist that their words be taken at face-value: the Society for Real Arithmetic insists that when they say 'there is an even prime' they mean that there is *really, out there* an even prime. I take it that members of such a community are speaking falsely, if it turns out that mathematical entities such as numbers do not exist. But despite this, the philosophical significance of the paraphrase for members of the population seems unchanged. The key point is this: whether or not a paraphrase captures the semantic content of language depends on all sorts of psychological and sociological contingencies that are simply irrelevant to the philosophical enterprise at hand.

(Two weakening moves might be attempted. First, to say that it is not semantic content that is at issue, but *pragmatic content* or questions of *what norms assertion*. However, I think that it is equally implausible to take the words of the Society for Real Mathematics to be non-committal at the level of pragmatic content or norms. Second, one might say the goal is not to describe linguistic practice *as it is* (the *hermeneutic* project), but to describe *how it should be* (the *revolutionary* project).[2] But now different irrelevancies come into view. The surrogate discourse may suffer from all sorts of practical deficiencies, when compared to the original. This fact is surely irrelevant to the philosophical interest of the paraphrase, but it does undermine a straightforward understanding of the paraphrase as giving a picture of *how the discourse should be revised*.[3])

My view of the goal of paraphrase is close to Quine's:

> likeness of meaning is not my aim.... One can still hold that [the surrogate] serves any purpose of [the original] that are worth serving.

[1] Paraphrase as semantic analysis can be understood within a model-theoretic approach to semantics in the way described by Lewis (1970a, p.205)

[2] The terminology is from Burgess and Rosen (1997). Compare Stanley (2001).

[3] Compare Melia (1995). There may be other non-revisionary readings of the 'ought' involved here with which I have no problem. Suppose, for example, it is glossed at what one with philosophical qualms *ought* to take from an assertion i.e. how the discourse 'ought to be understood' (compare Nolan (2002, §1.3), who uses both this gloss and the one I object to above). I expect that cashing out the force of this 'ought' will take us into issues of 'facts underlying the discourse' very much like those discussed below.

Quine (1960, p.214)

Unlike Quine, a scepticism about objective semantic analysis or synonymy in general forms no part of my motivation; but with Quine, I hold for the reasons sketched above that semantic analysis is irrelevant to the purposes of a metaphysical assay. If we label the earlier picture *paraphrase as semantic analysis*, we can call the Quinean view *paraphrase as regimentation*

What constrains reductive paraphrase, if not semantic equivalence? The basic idea is that the paraphrase reveals what the *facts underlying* the discourse in question are. This might strike some as more metaphor than substance. However, I think substance is introduced by the kind of consideration adduced a number of times above: we are looking for the facts that *play the central theoretical role* that the discourse is meant to play. Thus, if nominalistic paraphrases can do all the work that Platonistic arithmetical facts were supposed to, then (absent such Platonistic facts) we are justified in saying that what was *really going on*, what *underlay Platonistic talk*, what *Platonistic mathematics latched onto* or *was tracking* were facts expressed by the nominalist paraphrases.[4]

We can see why paraphrase that captures the 'facts underlying' a discourse in this particular sense will be philosophically interesting. One canonical mode of argument for believing in abstracta, possibilia or other problematic metaphysics, is to point to a certain role that they play (say the use of abstracta within physical science,[5] or the use of possible worlds within philosophy, linguistics, probability theory and so on[6]). If (1) the theoretical role cannot be achieved by other means, and we know that (2) it must be achieved *somehow*, then we have an argument for their existence. A reductive paraphrase in our sense debunks this argument, by undermining (1).[7]

Theory-shadowing paraphrases

Consider philosophical *fictionalisms*. Some paradigmatic examples include:

- Field (1980) on mathematics
- van Fraassen (1989b) on unobservables
- Rosen (1990) on modality
- Dorr (2002) on mereology

There is a certain core to these fictionalist proposals: a distinctive form of paraphrase. For example, for Field, the claim that there is an even prime corresponds to: *it is a consequence of standard arithmetic*

[4] I take it that the link between these metaphors and the discharging of a given theoretical role is relative to the philosophical purposes at hand. In other contexts, we might think of the facts that play a distinctive causal role in the genesis of opinions expressed in the discourse, as what 'underlies'. Consider, for example, the relation between physical properties of objects and opinions about what colour objects are, for the error-theorist of colours (Mackie, 1976). Quine (1960, §33) also emphasizes the context-relativity of successful regimentation/paraphrase.

[5] See Field (1980) and Colyvan (2003), and the references therein.

[6] See Lewis (1986c, §§1.2-1.5).

[7] I think that this conception of the utility of a fictionalist paraphrase can be found in Nolan and O'Leary-Hawthorne (1996, p.28):

> ...[mathematical fictionalist proposals] are a way of introducing a loose way of using talk about numbers that is not ontologically committing. [They] can say 'When I am speaking loosely, were I to say "The number of moons of Mars is identical to Two", take that as equivalent to "According to the fiction, the number of moons of Mars is identical to Two"...
>
> So long as the ...rules *give talk about numbers a useful role to play, they will serve their purpose*

(my emphasis.) Nolan and Hawthorne's notion of truth 'loosely speaking' is explicitly not to be read as a semantic proposal. It looks very close to my notion of 'supported by the underlying facts'.

that there are even primes. For Van Fraassan, what underlies *atoms contain protons* will be the fact that: *according to empirically adequate scientific theory, atoms contain protons*. For Rosen, a modal claim that *possibly, there are blue swans* means no more than: *according to the fiction of the plurality of worlds, there is a world containing blue swans*. Dorr holds that if I say that *there is a chair in front of me*, what I communicate is that: *Were there to be composite objects, there would be a chair in front of me*. The form of all these paraphrases is roughly as follows:

$\ulcorner p \urcorner \rightarrowtail \ulcorner$According to theory T, $p\urcorner$

Different fictionalisms then take different views about how to understanding 'according to', how to pick out the appropriate theory, and so forth.[8]

Beyond this core, fictionalisms are also characterized by (some or all) of a cluster of additional theses:

1. Statements in question are (often, and unexpectedly) strictly and literally speaking, false.[9]
2. Attitudes towards the contents in the area are not straightforward belief, but rather 'pretended belief'
3. Fictionalist surrogates specify the semantic/pragmatic content of statements in the discourse.[10]

For our purposes, it is best to divorce the core paraphrase proposal from any accompanying theses: this fits with our earlier discussion to concentrate on paraphrases that 'regiment away' metaphysical commitments, rather than getting involved in the contingencies of the fine character of the discourse in question. To avoid terminological disputes, however, we shall reserve the term 'fictionalism' for those more ambitious theories that aim to satisfy the cluster (1-3). Let us call regimenting paraphrases of the distinctive form "theory shadowing paraphrases". Theory shadowing paraphrases that are offered without commitment to the kind of semantic theses associated with fictionalism we call *quasi-fictionalist* accounts.

A number of theories that are not comfortably described as fictionalisms are quasi-fictionalisms in this sense. Some examples, along with their respective paraphrase proposals, include:

- Sider (2002) on possibilia
  $\ulcorner p \urcorner \rightarrowtail \ulcorner$'$p$' is entailed by every pluriverse sentence[11]$\urcorner$
- Eliminative structuralism (if-thenism) concerning arithmetic (Parsons, 1990)
  $\ulcorner p \urcorner \rightarrowtail \ulcorner \forall \bar{X}(PA(\bar{X}) \rightarrow p(\bar{X}))^{12}\urcorner$

[8] The views whereby p is a 'loose way of talking' about it's fictionalist paraphrase is called by Yablo (2001) a 'metafictional' fictionalism. I discuss the relation between metafictional fictionalism and Yablo's object-fictionalisms (when divorced from the semantic concerns that form a large part of Yablo's discussion) briefly in the conclusion to this chapter .

[9] In some versions, we replace the requirement that they be false when taken at face value, with the contention that language-users are in no position to know whether or not they are false. Thus the agnosticism of van Fraassen (1989b).

[10] We can make further distinctions: we can contrast 'semantic' fictionalisms that aim to specify the semantic content of our talk (compare Nolan's 'strong fictionalisms' (Nolan, 1997a)), with 'error-theoretic fictionalisms' where the problematic committal statements are literally false, but where the fictionalist paraphrase specifies some underlying pragmatic content to the speech acts in question. What Nolan (1997a) (following Rosen (1990)) calls 'timid' fictionalisms I do not think are fictionalist proposals at all in the sense at issue here.

How is semantic fictionalism compatible with (1) above? In such a case, we should read 'strictly and literally speaking' as an operator that *undoes* the work of the fictionalist's 'according to the fiction' operator. I.e. 'read at face value' they are false, but according to the present value, the semantic content is not given by a face-value reading. For discussion of this, and a possible relation between semantic and error-theoretic fictionalism, see Dorr (2002).

Yablo (2001) contains a nuanced discussion of possible semantic views a fictionalist might adopt.

[11] A pluriverse sentence is a certain kind of maximal description of the modal realm, which Sider proposes to analyze using *de dicto* modal resources.

[12] where PA is a second order formulation of the Peano axioms, wherein both PA and p all non-logical symbols replaced by variables.

- Hellman (1989) on modal structuralism
  $\ulcorner p \urcorner \rightarrowtail \ulcorner \Box \forall \bar{X}(PA(\bar{X}) \rightarrow p(\bar{X}))^{13} \urcorner$

Summary

Fictionalist views have a variety of characteristics. Fictionalisms are often associated with the view that, strictly construed, statements in the problematic area are often false, even if correctly assertible; and that the appropriate attitude to take to them is not belief, but rather a non-committal kind of acceptance. They assign a central role to a certain kind of paraphrase: associating a 'problematic' sentence S with the fictional paraphrase 'According to the fiction F, S'. *Quasi-fictionalism*, as I have introduced the term, drops commitment to everything but the central role for fictionalist paraphrase.

Quasi-fictionalism seems a fall-back position for any fictionalism. In the kind of example given above, where Platonist mathematicians demand that their words be taken at face-value, what might be put forward as a hermeneutic fictionalism for a less opinionated population becomes a reductive quasi-fictionalism. From the metaphysical point of view, the additional trappings that make a quasi-fictionalism into a fictionalism, are irrelevant in such cases. The significant feature is that the facts underlying the discourse in question (i.e. discharging the function that was the *point* of the discourse in the first place) can be outlined in metaphysically less committal terms.

[13] where *PA* is a second order formulation of the Peano axioms, wherein both *PA* and p all non-logical symbols are replaced by variables. $\Box$ is supposed to be 'logical necessity'—but clearly variant doctrines can be derived by interpreting this as epistemological, metaphysical or some other kind of necessity.

2.2 Incomplete fictions and theories

There are certain issues that impact all theory-shadowing paraphrases just in virtue of their distinctive form. One cluster of issues are technical questions about how the machinery of the paraphrase is to be understood . In the examples just used, we have seen a variety of ways of cashing-out 'according to' or 'follows from': as logical consequence relations, as universalized conditionals, as necessitated universalized conditionals, as counterfactual conditionals, or as operators drawn from an independently motivated analysis of locutions such as 'According to the Sherlock Holmes stories ... '. Clearly, how this notion is cashed is of key strategic significance for a reductive project.

There are also issues about the extent and tenability of the projects. Two of the most important are the following:

1. The issue of *Reflexive* application: what happens when we try to apply the paraphrase to sentences involving the problematic notions themselves. For example, what do we get if we ask about the modal status of claims about possible worlds, or about the number of numbers.

2. The issue of *Incomplete Fictions*: what happens when the theory involved is silent about some salient matter?

The latter, in particular, will be relevant to us when we come to consider how the interpretationist should understand inscrutability arguments, and so it is this that we discuss here.[14]

The puzzle of incomplete fictions

The puzzle of incomplete fictions is instructive, in that it illustrates how our conception of the paraphrase project can give a motivated resolution of a tricky issue. It will also be of strategic significance, since, on the view that I will be urging, inscrutability of reference may turn out to be a particular instance of this kind of puzzle.

The puzzle with which we are concerned arises when the fiction/theory on which our fictionalist paraphrase is based is *incomplete* in certain respects. The puzzle is tightest when we consider a fictionalism with semantic pretensions: the paraphrase aims to uncover what 'was really being said' all along. Consider a simple 'hermeneutic' fictionalism concerning set theory. This has it that claims such as 'there is a null set' corresponds to the claim that 'According to set theory there exists a null set'. Writing S for a formulation of set theory, the semantic pretensions of the view being considered lead us to endorse a 'fictionalist biconditional':

$$p \leftrightarrow \text{According to S, } p$$

Now it is well-known that there are sentences that are not only unsettled by the current axioms of set theory, but which look like in principle they cannot be settled. One such is the *generalized continuum hypothesis* which we write σ. If, as seems plausible, the set theoretic story is silent about this claim, we get the following pair of claims:

SET THEORETICAL INCOMPLETENESS

- $\neg$According to S, σ

[14] Problems for modal fictionalism arising from reflexive application are discussed in Brock (1993); Rosen (1993). For the generalization and discussion of the case of arithmetic, see Nolan and O'Leary-Hawthorne (1996). The line of thought suggested by Nolan and Hawthorne in that paper has much in common with the response to the incompleteness problem suggested below: they claim that the reflexivity problems can be avoided if we do not buy into fictionalism as a semantic analysis of the discourse in question. For a direct response to Nolan and Hawthorne, see Yablo (2001).

- $\neg$According to S, $\neg\sigma$

We can now derive both $\neg\sigma$ and $\neg\neg\sigma$, in each case by a single application of *modus tollens* on the fictionalist biconditional.[15]

The problem is clearly a general one, arising from (a) the endorsal of the fictionalist biconditional; and (b) an incomplete fiction/theory within the fictionalist paraphrase, leading to the pair of denials of the form given above.[16]

The point has received occasional discussion in the context of *modal fictionalism* (cf. Rosen, 1990). There, an incompleteness within the 'hypothesis of the plurality of worlds' is used to cause trouble for the fictionalist biconditional:

$$p \leftrightarrow \text{According to PW, } p^*$$

where PW is a story about the plurality of worlds as described by Lewis (1986c), p is a modal sentence, and p^* is a paraphrase of p into possibilist language (following Lewis (1968)).[17] Rosen's suggestion in the modal case is to reject the analogue of SET THEORETICAL INCOMPLETENESS. He recommends that we should formulate the fictionalist biconditional using an operator 'According to PW, p' which is neither true nor false (i.e., is indeterminate) wherever p is a sentence about which *PW* is silent. Now, many find this implausible if presented as a *general* account of locutions such as 'according to the fiction'; so the suggestion is often taken to be that we are to treat 'According to PW' as a *sui generis* primitive of the account.

One problem with this kind of tactic is the sheer implausibility of treating 'According to PW' as having no internal structure. Another concerns the costs of having to postulate a new *sui generis* piece of ideology; after all, it was reduction of *ideology* (e.g. 'Necessarily') that we were after in the first place. The primitivist strategy looks even less attractive from the wider perspective. If we were to adopt Rosen's response for each fictionalism threatened by the problem of incomplete fictions, then we would have in each case a new sui generis 'According to F' primitive. At the least, we want to discern structure within the operators and to give an independent account of their common core, 'According to F'. Multiplication of duplicate primitives is a sure sign we have gone wrong. I take it that since the puzzle of incomplete fictions is a *general* issue for fictionalisms, we should be looking for a *generally* applicable answer: and this is not what Rosen offers us.

(I think, in fact, that in the particular case he is interested in, Rosen can do better to defend his 'indeterminacy'-strategy. I see two lines of defence. First, Rosen's fictionalist could maintain that 'According to' is a general fictional operator. The burden would then be to explain away intuitions concerning, for example:

> it's not the case that according to the Sherlock Holmes story, Holmes' mother was born on a Tuesday.

In light of Conan Doyle's silence on such matters, this statement certainly seems to express a truth. However, generally constructions involving 'it is not the case that' attached to a less controversially

[15] Two caveats: first, it takes an extra move to obtain an *explicit contradiction* from this, i.e. to get the conjunction $\neg\sigma \wedge \neg\neg\sigma$. This move may not be as trivial as it first looks, as the setting has formal analogies to the *subvaluationist* settings where exactly this kind of move is denied, and a *paraconsistent* situation arises. Second, one might wonder whether the pair of statement really contradict each other: that is, do they express contradictory *contents*? The thought would be that, since their contents are given by the fictionalist paraphrases, which do not contradict each other, we have merely the superficial appearance of contradiction, not a genuine paradox.

[16] Note the parallel with well-known arguments for contradiction from denials of bivalence. See, for example, Williamson (1994, ch.5.). We have a contradiction where we have a 'disappearing' operator $O(p) \leftrightarrow p$ together with a denial of exhaustiveness $\neg O(q)$, $\neg O(\neg q)$. With O as truth, we get an argument against denials of bivalence. With O read as 'According to F' we get the problem of incomplete fictions.

[17] The specific incompleteness appealed to concerns the maximum cardinality of non-overlapping objects there could be.

indeterminate sentence can seem true. Certainly if Alfie is borderline bald, accepting ‘it is not the case that Alfie is bald’ seems tempting.

The second tactic is to find some non-fictional model for the operator. Consider, for example, counterfactuals as understood by Stalnaker (1984). When the antecedent A of a counterfactual is compatible with either of two situations, neither of which has a better claim to be ‘selected’ than the other, on Stalnaker’s analysis it is *indeterminate* which of these is selected, and if the scenarios disagree over whether B, it will turn out indeterminate whether $A > B$ is true. The suggestion would then be to read the fictionalist paraphrase as ‘Were the hypothesis of the plurality of worlds to obtain, then p’. Where PW is not explicit over whether B, either B situations are nevertheless determinately selected (in which case PW is not in the relevant sense silent over whether B) or else B- and $\neg B$-situations are equal candidates to be ‘selected’, and (by Stalnaker’s lights) the counterfactual will be indeterminate, just as Rosen proposes. (I make no claim that Rosen’s fictionalist will want to buy into this counterfactual analysis of her operator, just that in doing so she would have a natural backup for Rosen’s response to the incompleteness worry.))

There is a more detailed reason why the general strategy that Rosen likes is unlikely to generalize to the set-theoretic case we started out with, and the other fictionalist paraphrases in which we are interested. Consider again the simple set-theoretic version fictionalism outlined above. The theorist putting this forward will typically be *opinionated* about the way she would understand the operator ‘According to S’.[18] Within the philosophy of mathematics, for example, one is likely to find it cashed-out (a) in terms of logical consequence (either taken as primitive, or analyzed model-theoretically); or (b) in terms of the truth of a universally conditionalized statement; or (b) in terms of the truth of necessitated universally conditionalized statement.[19] Suppose, for example, we adopted the first view. We would have a biconditional of the form:

$$p \leftrightarrow S \Vdash \ulcorner p \urcorner$$

The following seems then to be a *metatheorem* of set theory:[20]

[18] One could always declare this to be a mistake. For example, we might think that fictionalisms should *always* be cashed-out in terms of counterfactuals (and exploit the Stalnakerian treatment to deal with the puzzle of incomplete fictions, as described above). For defence of such counterfactual imperialism, see (Dorr, 2005)

[19] See, respectively, ‘if-thenist’ formalism, Parsons (1990) on eliminativist structuralism, and Hellman (1989).

A little more care needs to be taken in formulating incompleteness in such cases. For example, in the case of eliminativist structuralism, the corresponding case is one where neither $\forall X\Omega(X) \rightarrow p(X))$ nor $\forall X\Omega(X) \rightarrow \neg p(X))$ is true. If Ω was a second-order formulation of ZFC, then GCH is a good candidate for such an incompleteness, as are large cardinal axioms.

[20] If the consequence relation involved is spelled out in the standard way, this will indeed be a metatheorem (Gödel, 1938; Cohen, 1963, 1964) . There are worries over the standard exposition, however. (A) the model-theoretic consequence relation is normally cashed out in terms of set-theoretical constructions. By Cantor’s paradox, there is no set of all sets, and hence there is no set-theoretical model whose domain contains all sets. Hence, it seems, no such construction can capture an ‘intended’ interpretation of set-theory, with a domain consisting of all the sets there are. (There are, however, results that give us reassurance that the standard treatment of consequence will not mislead. See Kreisel (1967). Moreover, there are those who reject the picture of a single ‘intended’ interpretation of set theory. See Zermelo (1930) and Parsons (1974).). (B) in the context of a fictionalism over set theory, a set theoretical explication of the key notion in the paraphrase is odd. One way out of this is to use a formulation within type theory of the kind suggested by Shapiro (1991) and defended by Rayo and Uzquiano (1999) under a plural interpretation of the style promoted by Boolos (1984). Another is, with Field (1989), to take the notion of consequence as primitive. (C) The focus on consequence, as opposed to conditionals, allows us to avoid some of the issues concerning the finite expressibility of theories. In the variants where conditionals are appealed to, a higher-order setting becomes attractive, enabling, for example, a finitely axiomatization of Peano Arithmetic. But a higher-order setting makes controversial the kind of incompleteness claims to which we are appealing.

There is a question whether incompleteness results go through under these interpretations. For in general, a more powerful notion of consequence (in connection, for example, with a higher-order formulation of the theory at hand) may mean that many more sentences are semantic consequences of the theory than a first-order understanding of consequence would allow. Our choice of example—the generalized continuum hypothesis—is fairly robust. For a second-order formulation of set theory is only *quasi-categorical*—it leaves unsettled ‘how tall’ the hierarchy of sets might be. Whether or not the generalized continuum hypothesis holds might well depend on the height.

In general, I am inclined to think that the puzzle arises in other cases, even if we move higher-order, for I think that the natural

SET THEORETICAL INCOMPLETENESS 2

- it is not the case that $S \Vdash \ulcorner \sigma \urcorner$
- it is not the case that $S \Vdash \ulcorner \neg\sigma \urcorner$

With set-theoretic fictionalism cashed out in this way, there is no room for flexibility on incompleteness: we must look elsewhere to resolve our puzzle. The general *methodological* point is that questioning the incompleteness premisses of the puzzle of incomplete fictions will only be applicable where the operator in terms of which the fictionalism is phrased has the (*prima facie* odd) features which Rosen postulates in the modal case. In many cases, then, Rosen-style responses will be simply inapplicable. Thus, we are motivated to look for a different resolution.

Incomplete paraphrases

I think that the key to defusing the puzzle above lies in the conception of paraphrase we discussed above. There, it was argued that reductive paraphrase is first and foremost a project of describing the *facts underlying* a discourse. Philosophical significance is not carried by the contingencies of what our expressions actually mean or commit us to, but what commitments we *have* to undertaken in order to discharge a certain theoretical role. This view enjoins a *neutrality* about the semantics of the discourse in question. It is compatible with an error-theoretic approach, whereby the absence of ontology renders the committal claims of the discourse false. It is also compatible with the view that the semantics of the given domain is interestingly related to the paraphrases in question.

Now, if we drop the claimed semantic equivalence of original statements and fictionalist paraphrase, we lose our motivation for the fictionalist biconditional. This by itself is enough to block the derivation of paradox. However, it is not so easy to escape the puzzle. To begin with, it resurfaces, albeit in a watered down form. We have a pair of contradictory sentences, σ and $\neg\sigma$, that are each paraphrased to something false. That still looks uncomfortable. Secondly, if we are to maintain *neutrality* over semantic standing of the discourse, we need to block the derivation of conditional *under the assumption that the paraphrase discloses the semantic content of the original discourse*. Without something further to say, it looks like we have an argument for an error-theoretic approach.

Biting those bullets, and buying into an error theory, might in the end be the appropriate response. I think we can do something better, however. Consider again the conception of reductive paraphrase outlined previously. The basic idea of theory-shadowing paraphrase is that the facts underlying a discourse D are facts about what follows from a certain fiction or theory F. The key constraint on the paraphrase is that the surrogate sentences discharge the theoretical role appropriate to the original discourse.

Why then do we assume that a paraphrase must be *total*, that it pair *each* sentence of the original domain with a surrogate? If some of those sentences do no work (i.e. if we never, in discharging the theoretical roles appropriate to the discourse, appeal to them) then we have no *need* for a paraphrase. Nor does the theory-shadowing conception motivate any paraphrase in cases where F is silent: it is facts about what *does* follow from the fiction that are supposed to underlie the original discourse, not facts about what *does not* follow. My suggestion, therefore, is that we take the reductive paraphrase to be partial, to take the form:

$$\ulcorner p \urcorner \rightarrowtail \ulcorner \text{According to F, } p \urcorner$$

way to understand higher-order consequence is through Henkin-semantics, which do not allow us to establish the categoricity results that would undermine completeness claims (see the second section of Appendix B. This is a matter for another occasion.

for exactly those cases where the disjunction '(According to F, p) or (According to F, $\neg p$)' holds. This is all that is motivated by the theory-shadowing conception, and all that is needed for it to discharge its functions.[21]

Since the paraphrase is partial, we eliminate any analogue of the incomplete fictions puzzle. What we get instead is the general statement that neither p nor $\neg p$ is paired with any 'underlying facts'. That gives content to the claim that there *is no fact of the matter* over such p. We can give a formal elucidation of this notion. For the general case of a theory-shadowing paraphrase based on 'fiction' F, let us introduce a new operator FACT, *defined* through the following biconditional:[22]

FACT[p] iff According to F, p^*

We can now, in a neutral way, state the situation at hand. Where σ is the generalized continuum hypothesis, we have:

$$\neg\text{FACT}[\sigma] \wedge \neg\text{FACT}[\neg\sigma]$$

Notice that the disjunction FACT$[\sigma \vee \neg\sigma]$ does hold: in the jargon, the facts are 'non-prime'.

We noted above that our conception of paraphrase motivated *neutrality* on the standing of the biconditional, not its rejection. What we now show is how a semantic hypothesis could be maintained in the setting just described. Imagine that we supplement the core theory-shadowing paraphrase, outlined earlier, by the *additional* view that, the fictionalist paraphrase gives the content of our set-theoretical talk. We regard the apparently committal talk as a 'manner of speaking', perhaps convenient for practical purposes, when what's really being expressed is given in the fictionalist paraphrase. *Given* this additional principle, we now get the biconditional back. Where 'p' is reductively paraphrased to 'According to F, p', we have:

$$p \leftrightarrow \text{According to F, } p^*$$

But notice that there is an absolutely natural restriction on this scheme: it holds *only where the paraphrase relation obtains*. According to the view given above, it *will not* hold in cases where the modal fiction falls silent. Thus, we derive the result that Nolan (2002) urges: rather than fiddle with the behaviour of 'According to' to deny INCOMPLETENESS, we restrict the biconditional.

Nolan further claims that modal claims not covered by this should be taken as indeterminate, thus getting behaviour matching Rosen's original hypothesis. Can we derive this claim also? Not directly, I think. As we imagine the case, we have a community who explicitly decide to let their set-theoretic or modal talk be in effect a round-about way of talking about the fictionalist paraphrases. At the points where the paraphrase is silent, we simply do not express anything, however round-aboutly: their convention does not cover the case. It might be that one takes situations where linguistic conventions lapse, to be ones where the relevant sentences have an indeterminate status: neither true nor false. If so, then one has indeed derived the second aspect of Nolan's suggestion. [23]

[21] Of course, it would be a good objection to this view if it could be shown that some p over which the theory is silent does do some important work. This would be a problem for *any* theory-shadowing proposal based on that fiction, not a problem specific to the case at hand.

[22] Compare Fine (2003).

[23] The current proposal, though compatible with Nolan's, differs in two respects. First, it gives a non-*ad hoc* motivation for the obtaining of the biconditional, as the joint upshot of (a) the project of paraphrase; (b) a linguistic convention to let one's modal discourse express their fictionalist paraphrases. Second, it is a response that gives a *general* recipe for treating incomplete fictions. Dropping (b) leaves the fundamental diagnosis of the situation untouched: thus we can apply our result theorists in an *error theoretic* treatment of modal discourse, underpinned by the possibility of fictionalist paraphrase.

2.3 Interpretationism as a quasi-fiction

Recall that the puzzle was to see what the metaphysical view associated with interpretationism was. I will here draw on the resources examined above to give one gloss on the enterprise. My suggestion is that interpretationism is a theory-shadowing paraphrase put forward in furtherance of the project of reducing intentional idioms to non-intentional ones. That it shares this general shape is suggested by the gloss given earlier ‘for an expression to have a semantic property is for that property to be assigned to it by the best semantic theory’. (Clearly, such a principle is most comfortably read in furtherance of a *reductive* rather than *semantic* project—our aim is to *found* semantic properties, not to give the semantics for talk about semantic properties.)

The rough characterization of the position, then, can be given by the following scheme:

$$\ulcorner p \urcorner \rightarrowtail$$
$$\ulcorner \text{According to best semantic theory } p \urcorner$$

There will be immediate concerns about how this can really work as an ideological reduction of the semantic. If “Londres’ refers to London’ is to be replaced with ‘According to best theory, ‘Londres’ refers to London’, we seem to *use* the ideology we were meant to eliminate (‘refers’) within the surrogate. Moreover, one might be concerned whether ‘According to best theory’ is to be understood: it sounds very like the frankly semantic ‘Best theory says…’. (This concern is the analogue of the worries expressed by Nolan (2002, §4.2) concerning ‘strong’ modal fictionalisms: that it appears that they will make circular appeal to a modal notion of *entailment*.)

The concerns are addressed by further details about how the machinery of the theory shadowing paraphrase is to be cashed-out in this particular case. We avoid *using* ‘refers’ in the surrogate sentence, by *mentioning* it instead. The paraphrase really pairs “Londres’ refers to London’ with ‘⌜‘Londres’ refers to London⌝ follows from best theory’. The concern about possible circularity induced by appeal to “according to” or “following from” are outlined below and answered in the following section: essentially, the key is to analyze ‘follows from’ in terms of (something like) logical consequence, and argue that no residual circularity remains.

A second immediate objection rests on the observation that *semantic* theory, per se, will not entail *categorical* semantic truths such as

> ‘London is a large city’ is true.

At best, we get:

> ‘London is a large city’ is true if and only if London is a large city.[24]

Appeal to non-semantic facts—facts about the size of cities—is required to discharge the right-hand side and allow us to derive the canonical formulation. Recall that the biconditional is not the underlying story—it may not even be true. What is really going on is:

> According to best theory, [‘London is a large city’ is true if and only if London is a large city].

and we can’t derive the categorical statement from this together with *London is a large city*.

[24] Or, in the possible-worlds setting: ‘London is a large city’ is true relative to all and only worlds where London is a large city.

This is a general issue with fictionalist-style paraphrases. In the case of modal fictionalism, it is mentioned by Rosen (1990) (see also the discussion in Sider (2002)).[25] I take the appropriate response here to be the same as in that case. Really, the paraphrase should not be to 'According to best semantic theory, *p*', but to 'According to best theory *together with the non-semantic facts*, *p*'. By incorporating an 'encyclopedia' of non-semantic facts into the theory that our paraphrases shadow, we can allow categorical statements such as that cited above to be derived.

Our paraphrase, then, will quantify over two kinds of theories. First, there will be the *selected* semantic theory: in Lewis' terminology, that grammar which is *used* by the community in question. It is the substantive work of interpretationist metasemantic theory to say what this is. Second, there will be the *encyclopedia*—a theory encapsulating all the non-semantic facts. Putting together these last remarks with the earlier ones about the use of quotation and model theoretic consequence, we get a more precise (though still schematic) statement of the interpretationist paraphrase:

$$\ulcorner p \urcorner \rightarrowtail \ulcorner \forall x, y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash \text{`}p\text{'}) \urcorner^{26}$$

This, then, is the form of the proposal. The fundamental question now is whether all the machinery here invoked—the predicates SELECTED and ENCYCLOPEDIA, the consequence relation $\Vdash$, and the use of quotation—can each be explicated without appeal to semantic facts. We shall outline four areas of concern.

Selected theory

Recall the characteristic two-step approach to identifying the semantic theory appropriate to a population—we first pick out data (e.g. conventions of truthfulness, or T-sentences) and then let the correct semantic theory be the best one that explains the data. We can regard such stories as implicitly constructing the required predicate 'SELECTED'. For we can say that a theory is selected iff it is best at explaining relevant data, where this data takes the form (for example) of conventions of truthfulness in population *P*. Semantic ideology, then, is reduced to whatever ideology is used in explicating SELECTED. Attention to the details of the interpretationist metasemantic account thus allows us to assess its success as part of a reductive paraphrase. Characteristically the interpretationist paraphrase will feature a predicate DATA—holding of the data the interpretationist identifies, and a relation BEST, holding between this data and the semantic theory that best accounts for it.

$$\ulcorner p \urcorner \rightarrowtail \ulcorner \forall z, x, y(\textsc{data}(z) \wedge \textsc{best}(z, x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash \text{`}p\text{'}) \urcorner$$

Here different interpretationisms diverge. Appeal to mental representation will be made explicit in the way the Lewisian unpacks DATA; whereas the characteristic Davidsonian appeal to 'holding true' will feature in the Davidsonian's explication of the same relation.

The other component to picking out the selected semantic theory is equally vital: What is it that makes a semantic theory the 'best' account of this data? Again, if appeal to intentional or semantic

[25] Rosen's fictionalism wants to earn the right to assert a categorical modal claim such as "Possibly, there is a swan." The basic fictionalist proposal pairs this with 'According to the modal fiction, there is a possible world where there is a swan'. All that follows from 'the hypothesis of the plurality of worlds' *per se* is the conditional that *if* swans exist in the actual world, then there is a possible world where swans exist. We need again to appeal to facts about the actual world to discharge the antecedent. Rosen therefore adds an appeal to an 'encyclopedia' of actual facts, just as we are about to.

[26] In fact, as explored below, the use of quotation marks hides some complexity: really, we should use an explicit "quotation function" $*p*$.

Note the *universal* quantification used in this formulation. An alternative would be to use existential quantification. The two are equivalent so long as on e and only one theory is selected—see Chapter 3.

notions must be made in spelling out this notion, that impacts on the reductive ambition that can be sustained by a given interpretationism. In fact, particular proposals for unpacking this notion will be the focus of later parts of this thesis. Our methodology shall be to start with a 'null' hypothesis: an explication of 'best theory for data-set z' ('BEST') that minimizes appeal to contentious resources.

The null hypothesis I have in mind I call BEST=FIT. This identifies the relevant sense of *best theory* of a certain set of data, with a theory that *fits* the data, in the sense that, ideally, it (1) entails the data and (2) entails nothing incompatible with the data.[27] (This will require modification if we are to deal with situations data are less than totally coherent, and so, intuitively, no theory 'fits the data' perfectly.[28])

Taking the second component of an interpretationist proposal to be *fit with the data* has a distinguished history—it is (explicitly or implicitly) endorsed by Davidson, Lewis and Putnam at various points.[29] BEST=FIT is a crucial premiss in the inscrutability puzzles which are the focus for much of this thesis. It is important, then, to note that we have offered no positive reason for endorsing it. On the other hand, any more constraining hypothesis will have the burden of showing that it is not appealing to any problematically semantic resources.

With the other three worries, I will list the problems here, and then develop the framework that resolves them in the following section.

The Encyclopedia

There are two concerns with the analysis of the predicate 'encyclopedia'. One is whether we can really have a *theory* which has the intended effect. What reason have we to think that any language has the expressive resources to capture *all* the facts about the world? The other is how we should analyze this predicate. The natural way to go to describe a theory containing all the non-semantic facts would be talk about a *true* theory whose *content* is wholly non-semantic. Neither notion, of course, can legitimately be appealed to at this stage.

Consequence

In order to eliminate suspicion that semantic resources are smuggled into the notion of being true 'according to' some theory, we will appeal to the relation of logical consequence. This stands a chance of being characterized without appeal to the semantic.

What is this notion? One option is to understand it model-theoretically. The rough characterization is as follows: that ϕ is a model-theoretical consequence of Γ iff *no matter what the language means* we never get all of Γ being true (on that interpretation) but ϕ false. Despite the talk of 'truth' and 'interpretation', there is no blatant circularity. For given a syntactical characterization of the language, we can define what it is for a mathematical function to be an 'interpretation' of the language in a formal sense (i.e. just a mapping from names to objects, predicates to sets of objects, etc.) Further, we can then explicitly define the property of truth-on-interpretation-f (e.g. *PN* has this property if $f(N)$ is a member of $f(P)$). So far, so good.

One might worry about a more subtle circularity. In explicating consequence model-theoretically, it is no good letting *every* syntactically appropriate interpretation function be ranged over. We

[27] I ignore, for the time being, complications arising from the need to introduce pragmatic factors mediating between best semantic theory and the data.

[28] Progress towards gradable notions of 'fit' are developed in Braddon-Mitchell (2001), who appeals to work on lossy algorithms, Rothschild and Leuenberger (2005) who focus on accounts of 'truth-likeness' in the philosophy of science, and Elga (2004), who in the case of probabilistic theories develops a notion of the relative 'typicality' of data-sets.

[29] See, for example, Davidson (1977, 1979); Lewis (1975); Putnam (1980, 1981).

want ‘A’ to be a consequence of ‘$A \wedge B$’ in the relevant sense. However, if we allow ourselves to re-interpret ‘$\wedge$’ as *disjunction*, this will not be a case. Accordingly, the standard Tarskian analysis requires we hold fixed the actual interpretation of a certain range of expressions: the logical constants.[30] It appears, then, that model-theoretic consequence makes essential appeal to the identification and interpretation of logical constants within the object language. *Prima facie*, semantic notions are being invoked.

Quotation

The appeal to quotation, in the course of giving the interpretationist's paraphrase, may give pause. For our aim is to get to facts that are expressed using quoted and unquoted words: indeed, it might seem that quotation is exactly the inverse of the reference relation, so that given the former we could explicitly define the latter. To get a grip on the concern, suppose that we write the quotation function $**$. Then it looks as though we could define ‘reference’ by means of the following: w refers to o iff $w = *o*$. We need to address the worry, therefore, that in using quotation we are tacitly assuming object-language/meta-language connections that are semantic in nature.

The key thought in addressing all three questions is to distinguish carefully genuine appeal to notions of *reference* and *truth*, from appeal to ‘reference-on-I’ and ‘truth-on-I’ where we can give an *explicit* characterization of these ‘parochial’ relations. Whereas what ‘Billy’ refers to is a semantic issue, what ‘Billy’ stands in the R relation to, where we explicitly stipulate what the extension of the R relation is to be, involves no appeal to anything representational. As we shall see in the following section, appeal to such explicitly defined relations allows us to develop a ‘parochial’ notion of consequence, quotation, etc. which will serve our purposes.

[30] We get a range of notions of model-theoretic consequence if we hold other aspects of interpretation fixed. I ignore this for now.

2.4 Parochial consequence and paraphrase

Consider again the reductive paraphrase of semantic discourse that has been offered.

$$\ulcorner p \urcorner \rightarrowtail \ulcorner \forall x, y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash *p*) \urcorner$$

Notice that we are now mentioning a language *within* the paraphrase. As already highlighted, it is important to be able to appeal to some *prima facie* semantic properties of this language (e.g. to pick out its logical constants, and say what their intended interpretation is.) The challenge is to say how we do we do this without compromising the reductive ambition.

This section will:

(A) state exactly what is involved in the above paraphrase;

(B) describe a ‘parochial’ interpretation of a language, and show that such resources are legitimate within a reductive project;

(C) describe how such resources allow us to explicate the quotation function $\ulcorner * \ * \urcorner$, and to construct a notion of parochial consequence $\ulcorner \Vdash \urcorner$;

(D) show how a specific ‘meaning building’ language can be developed within the base language over which the quantifiers within the paraphrase range and which can be taken to be the language for which the parochial consequence relation is defined. This will allow us also to define explicitly the predicate ENCYCLOPEDIA.

(A) The paraphrase

The general schematic formulation of the paraphrase has as instances statements such as:

$$\ulcorner \text{‘London’ refers to London} \urcorner \rightarrowtail \ulcorner \forall x, y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash \text{“‘London’ refers to Lond}$$

The paraphrase is given in the theorist’s (meta)language but there are several other languages in play. Let us give them names:

- Call the language *used* in the left-hand side of the paraphrase the *problem language*.
- Call the language *mentioned* in the left hand side of the paraphrase the *target language*.
- Call the language *used* in the right-hand side of the paraphrase the *base language*
- Call the language *mentioned* by the base language the *meaning-building language*.

Typically, in giving a paraphrase we are trying to show how the work done by a problem language (in giving the semantic properties of the target language) can be done by a base language that makes no mention of semantic resources. Here, our base language itself contains ‘higher-order’ resources—the ability to talk about the meaning-building language.

In the case just given, all the languages were (fragments of) English, albeit supplemented with some formal vocabulary. The problem language is the semantic fragment of English, which mentions English expressions. The base language is a non-semantic fragment of English, with special predicates and the resources to specify *syntactically* the sentences of the semantic fragment of English—so English is the meaning building language too.

The free use of quotation and other devices cannot be supposed to be in goodstanding. To bring this out suppose that the meaning building language is *French*. Then our ‘quotation marks’ need to take an English expression ‘dog’ to the French *quoted* expression ‘$\langle\langle$chien$\rangle\rangle$’. Clearly, here some sort of translation is being implicitly appealed to: something that is obscured by the all-English version.

As anticipated above, this will turn out to be unproblematic. For the purpose of clarifying our commitments we shall assume that the meaning-building language is a formal language, rather than ordinary English. This will mean that we are obligated explicitly to define the quotation function **. That we can do this without the use of representational notions will vindicate the somewhat sloppy formulation where we use English and ordinary quotation.

(B) Parochial interpretations

The key idea is to *use* predicates of the base language to specify aspects of an ‘intended’ valuation of the meaning-building language. Suppose, for example, that our meaning building language is an initially uninterpreted first-order language, containing the non-logical predicate ‘Q’. Then we can *stipulate* that this symbol will have the valuation $\{x : x \text{ is a quark}\}$. Or better: we stipulatively define a notion ‘refers-in-the-meaning-building-language’ which relates the symbol ‘Q’ to the set $\{x : x \text{ is a quark}\}$. We incur ideological commitment to quarks, ontological commitment to sets, but no commitment to genuine representational properties for the meaning-building language.

Now the notion ‘refers-in-the-meaning-building-language’ has been explicitly defined. This is not to explicate a general notion of *reference*—the relation that we have introduced is simply undefined beyond the narrow range of predicates for which we introduce it. Call it, therefore, a *parochial* interpretation.

A base language can therefore construct a meaning building language as a purely mathematical entity. and specify a parochial valuation by *using* base-language predicates as appropriate. This is no more problematic than any other use of predicates within a mathematical construction. There are no hidden semantic notions here.[31]

*(C) Applications: ** and consequence*

We are not assuming that the meaning building language is identical to the problematic language. The quotation involved in the paraphrase is not a purely syntactic device—it needs to correlate a given sentence of the problematic language with an appropriate expression in the meaning building language which is then quoted.

The required translation is as follows. First, the vocabulary to be reduced (e.g. ‘refers’) is mapped to a special symbol of the meaning-building language, ‘VAL’, say. Otherwise, an expression ‘London’ is translated within the meaning building language with some symbol L which refers-in-the-meaning-building-language to London. Therefore, the translation can be set up simply by using (1) resources shared by the problematic language and the base language; (2) parochial interpretations of the meaning building language.

The other resource that we promised to explain is that of consequence. I favour a model-theoretic analysis of consequence—but the standard model-theoretic definition requires we identify *logical* vocabulary, to tell us what interpretations are and which aren’t admissible. How, when reducing semantic properties, can we appeal to this? The solution is not to appeal to it. We define a *parochial* notion of consequence for the meaning building language.

As before, we can pick out symbols within the meaning-building language e.g. $\ulcorner\wedge\urcorner$ and *use* the base language to specify a parochial interpretation for them. We then can define a notion of ‘admissible valuation’ by restricting the range of valuations to just those that agree with our parochial interpretations on the assignment to the range of expressions picked out. We therefore have a definition of ‘$\Vdash$’, given just

[31] If the base language contains a quotation device, then it can treat itself as a meaning building language. The required relations are then ‘disquotationally’ specified.

mathematics, the syntax of the meaning-building language and again the *use* of various base language logical notions. Notice that because we have not said what it is for an arbitrary symbol to be logical, this is not adequate as a characterization of logical consequence—but the model theoretic notion just characterized will serve our purposes.[32]

(D) Lagadonian languages

It is clear from the above discussion that any meaning-building language could serve our needs—English is used for ease of presentation, but to legitimize its use, we need to define a parochial interpretation, so we could as well have used a purely formal language. In fact, for some purposes, a formal language has advantages over English. Recall, for example, the need for an *encyclopedia* that contains all facts about the world. What guarantee do we have that any natural language could be expressively adequate for this?

To allay such concerns, we can take our meaning-building language to be *Lagadonian* language (Lewis, 1986c, pp.145-6). This (purely formal) language has some distinctive features:

- The formal syntax of a categorial grammar.[33]
- It has as constants all objects.
- It has as predicates all sets (pure or impure).[34]
- There is a special uninterpreted relation-symbol VAL (for 'semantic value').[35]

The Lagadonian language so constructed is a mathematical entity (a proper class of sets). What makes it useful is the ease with which we can introduce a parochial interpretation for it. We can reductively analyze 'refers-in-Lagadonian':

> x refers to y iff $x = y$.[36]

We can therefore take this as a canonical framework in which to apply the treatment of quotation and consequence given above. Since we can now easily formulate a parochial notion of true-in-Lagadonian,[37] we can reductively define the encyclopedia predicate:

> ENCYCLOPEDIA(x) iff $\forall y$(y is true-in-Lagadonian$\rightarrow$ $y \in x$)[38]

[32] Even if we were to take logical consequence as a primitive, following Field (1989), we can make a similar case. Presumably something is a consequence of a set of sentences only *relative to* some interpretation of the logical constants involved (sentences, after all, do not have their consequences *essentially*). We can equally define a parochial notion of logical consequence as what the logical consequences of the formulae *would be* were the symbols to receive such-and-such an interpretation (explicitly specified).

[33] See Ajdukiewicz (1935); Lewis (1970a); Cresswell (1973). See also discussion in §5.1, below.

[34] Hence, some predicates will also be constants. These will be distinguished by syntactic position. If one likes, one can take a predicate to be the pair $\langle M, S \rangle$, where the first element is a name for the appropriate syntactic category (e.g. S/N) and the second element is a set of objects.

[35] We can introduce a special syntactic marker which will distinguish VAL from other relation symbols.

[36] This will not cover VAL, of course.

[37] Compare the definition of truth for structured propositions in Soames (1989). His 'structured propositions' (or 'structured intensions') are exactly Lagadonian sentences, from the present perspective, and the definition of truth for them is a parochial interpretation. Propositions, so characterized, are not intrinsically representational (compare Lewis, 1986c, p.146)(it is not clear why they are especially appropriate as truth-bearers, therefore, though clearly they are technically elegant).

[38] Notice that the base languages quantifiers have here to range over a proper class of entities. I shall assume that this raises no new problems.

2.5 Conclusion

The basic idea that meaning facts are ‘fixed by best theory’ connects interpretationism with a range of deflationary metaphysical accounts: on modality, mathematics, composite objects and in many other areas. The full ‘fictionalist’ package—including, maybe, an error theory about the semantic content of such claims, a ‘pretence’ account of attitude to such statements; or alternatively claims of semantic equivalence between original and surrogate statements—contains much about which we can and should remain neutral. The key idea that interpretationism takes from fictionalism is that of giving a ‘theory-shadowing paraphrase’ of the target discourse. In accordance with the general project of reductive paraphrase, this is intended to say what the facts underlying the discourse are, and is neutral on the question of the truth or falsity within the discourse itself.

In fact, there is more than one way in which we can specify the ‘facts underlying the discourse’ by exploiting the fictionalist paraphrase. Given a substantive account of how best semantic theory is selected, we can look at the *set of situations where all selected semantic theories entail S*. This is the recipe for constructing what Yablo (2001) (in analogous cases) calls the ‘real content’ of such sentences.[39] Given an interpretationist story about theory selection, the real content of a semantic claim in this sense would comprise worlds where (roughly) the patterns of assent and dissent are such as to select a semantic theory entailing the claim. In this way, we can ‘unwind’ our theory-shadowing paraphrase to associate a proposition (set of worlds) with a semantic statement. But thinking of Yablo’s ‘real content’ as the facts underlying semantic talk is not in competition with my claim that the theory-shadowing paraphrase gives these facts (individuated in a coarse-grained way, and supposing that words quoted exist necessarily): for the Yablo-real content holds in exactly the worlds where the theory-shadowing paraphrase $\forall x, y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash *p*)$ is true. Hence, the Yablo-real content and the fictionalist paraphrase are just ways of presenting the same underlying (coarse-grained) facts.

The distinctive interpretationist contribution to the theory-shadowing paraphrase of semantic facts is a particular style of account of what it is for a semantic theory to be *selected*. A semantic theory is selected if it is the best theory of the sentential data that the interpretationist identifies.

What happens if *more than one* semantic theory meets the conditions that the interpretationist sets down? That is, what if the constraints that the interpretationist sets down *underdetermine* best theory? There is here a tension within the account. The more constraints packed into the interpretationist’s account of what makes a semantic theory ‘best’, the more likely it is that resources used vitiate reductive ambitions. Conversely, the thinner that we make the notion of ‘best theory’, the greater the chances of underdetermination occurring. Does underdetermination arise? If so, how should it be handled? These issues are to be discussed in the following chapters.

[39] See §5 of Yablo (2001): “The real content is the circumstance K that *makes S* fictional” Yablo’s setting is slightly different from mine, as he is assuming a *particular* fiction is fixed, relative to which sentences are ‘fictional’ or not depending on how the world is. But since our theory-shadowing paraphrase *quantifies over* fictionalisms, we need to factor in this variation into the characterization of the real content: hence the version given above.

Part II

Arguing for inscrutability

Introduction to Part II

Discussion of inscrutability of reference will occupy the next three chapters. Within an interpretationist setting, inscrutability arises when there are *multiple* theories, each with equal claim to be the 'meaning-fixing semantic theory'. Chapter 3 will be concerned with the content of inscrutability *theses*: how to handle situations where we have a range of 'successful' theories by interpretationist lights, and what the content of claimed inscrutability is.

Chapters 4 and 5 present arguments for two types of inscrutability. The first concerns how reference of general terms such as "Rabbit" is *divided*: i.e. which objects fall under the predicate. The source here is Quine's famous "argument from below" (Quine, 1960).

The second type of inscrutability in which I will be interested is *radical inscrutability* of reference. This is the claim that there is no fact of the matter at all concerning which objects singular terms pick out (and indeed, which objects are in the extension of predicates).

In each case, I am interested in the first instance with how the interpretationist should view these arguments. I argue that we can treat these arguments as showing the costs of various forms of interpretationism—particularly of forms that endorse the attractively simple BEST=FIT hypothesis. Therefore, the issue to be discussed here is whether the semantic theories embedding strange ways of dividing reference, or radically permuted reference schemes, can fit with whatever data is thrown at them at the level of pairings of sentences with truth-values or propositions. It may well be that a different setting (e.g. causal theories of reference) would avoid the difficulties here sketched. That is not at issue for now.[1]

The two chapters present rather different cases for this, however. In the case of radical inscrutability of reference, we can produce formal theorems making out our results: we can prove theorems that appropriate settings will "overgenerate" semantic theories that are successful by interpretationist lights. In the division-inscrutability case, however, I will not be arguing for such a general result (indeed, I will be adducing counterexamples: instances of sentences which are true on one reading but not on another). My methodology will therefore be to present a semantic theory for a fragment of English, and show how in a range of *central* cases, the truth-conditions generated by rival ways of dividing reference are in each case "good enough".

If we manage, in either the division- or the radical-inscrutability case, to show that the relevant semantic theories match each other over the relevant data about the usage of whole sentences, interpretationists are then faced with a dilemma. Should they complicate their story about the foundations of semantics, denying BEST=FIT, for example, by putting further constraints on 'best theory'? Or should they instead learn to live with the inscrutability? I shall suggest that the division-inscrutability should not be too unpleasant. Radical inscrutability is another matter: and in the final chapters of the thesis we shall measure the costs of accepting it, and the burdens that would need to be undertaken in avoiding it.

[1] Such theories have their own problems to deal with. Some of these problems may impact on how they would respond to the ideas here presented, though: in particular, it looks like an account of what privileges a semantic theory dividing reference over Rabbit-slices rather than over Rabbit-worms would have to appeal to a solution to the "*qua*" problem (For a presentation of the *qua* problem, see Sterelny (1990, ch 6.)).

Chapter 3

Inscrutability theses

Our focus now is on the alleged phenomenon of *inscrutability of reference*. Such theses take the form of claims such as:

> There is no fact of the matter whether 'Rabbit' applies to instantaneously existing entities (rabbit-slices), or to perduring objects (rabbit-worms).[1]
>
> There is no fact of the matter about whether "Londres" refers to London, or whether it refers to New York.[2]
>
> There is no fact of the matter whether 'mass' denotes relativistic mass, or whether it denotes rest mass.[3]

We shall see in due course arguments for particular claims of this sort, and shall assess their effectiveness and tenability. Within interpretationism, we can expect arguments for inscrutability to take a particular form. What underlies the semantic, on this view, are truths of roughly the form 'It follows from the best semantic theory that *p*'. This raises the question: what if there is no unique best theory? What if there are a number of theories—even infinitely many—tied for 'first place'? Generally speaking, inscrutability results will threaten if we can make the case that semantic theories entailing the contradictory results fit the interpretationist's data equally well. [4] (For the moment, we will suppose that the only theories in the race to be 'best' are classical semantic theories.)

The task of this chapter is to gain a rigorous understanding about what it means to say that 'there is no fact of the matter...' over whether or not *p*. To this end we present two different ways of handling such locutions. The first is to construct from the multiple best theories an 'overall' meaning-fixing theory that encapsulates all of them. The second is to stick with the simple identification of selected theories with best theories, and allow there to be multiple selected theories. We discuss each of these in turn.

[1] See Quine (1960, ch.2.)

[2] See Davidson (1979).

[3] See Field (1973).

[4] Note that we *hold the metasemantic theory fixed* while considering whether inscrutability emerges. Thus we are here concerned with "genuine" inscrutability and not merely "conventionality of claims about reference" in the sense of Field (1975). However, see the discussion of 'innocuous' vs. 'illuminating' and 'surprising' inscrutability at 62, below.

3.1 A supervaluational treatment of inscrutability

Cashing out inscrutability of reference in terms of there being 'no fact of the matter' about which of a range of things a given term refers to, brings to mind other areas of the philosophy of language where such glosses have been offered. *Vague* language—paradigmatically predicates such as 'is a heap', 'is bald' and'is red'—lends itself to such elucidations. Faced with a series of men with receding hairlines, it is tempting to say that there is *no fact of the matter* about where the boundary between bald and non-bald men lies. Consider also counterfactual conditionals: in particular, the claim 'If I had flipped that (fair) coin, it would have come down heads'. Some maintain that given the chancy nature of the situation there is *no fact of the matter* whether this claim is true.

My ambition here is not to say anything novel about these claims (though the relationship between vagueness and inscrutability, in particular, is extremely interesting.[5]) What I intend to do in this section is to describe and elaborate one of the pieces of formal machinery that has been at the heart of debates about vagueness, and outline a possible analysis of inscrutability in terms of it. The machinery is commonly associated with *supervaluational* theories of vagueness. Classic papers on this include Lewis (1970a); Dummett (1975); Fine (1975); Kamp (1975). A similar setting is deployed by Field (1973, 1974) specifically in application to inscrutability. I will not trace the history of the notion here, but rather develop a version from scratch.[6]

In the following sections, I outline the basic idea of multiply intensional semantics by outlining two paradigmatic examples: relativization to possible worlds and to variable assignments. We see two illustrative 'derelativization' strategies: the picking out of a privileged index (e.g. the actual world); or quantification over all indices (in the case of variable assignments), and characterize local and global consequence in that general setting. Introducing a new parameter of 'delineations' gives the basic framework for a supervaluational treatment. (In Appendix A, I argue *contra* Williamson (1994, ch.5) and Keefe (2000, ch.8) that no matter whether we characterize consequence globally or locally in the context of supervaluational semantics, no revision of classical logic is induced.) I finish by defining various de-parameterized semantic notions that allow us to express inscrutability of reference within a supervaluationist treatment, and showing in each case how an object-language operator can be constructed to allow expression of inscrutability within the object-language.

Multiply indexed semantics

Following Lewis (1970a), we will treat indeterminacy within a model-theoretic semantics for a language. The model theory used, in general, will be *multiply intensional*, in the sense that expressions will be assigned extensions relative to a string of *indices*. Two familiar cases are as follows:

(A) One of the indices is a *possible world*. Thus 'Britain is in the Arctic' will be assigned a function that will map to *True* situations where Britain occupies that more northerly position; other situations (such as the actual one) will be mapped to *False*. We can then treat modal operators such as 'Necessarily' through clauses exploiting this parameter: 'Necessarily, Britain is in the arctic' is true at a world iff 'Britain is in the arctic' is true at every world.

(B) Another of the indices is a *variable assignment*. For example 'x is male' is assigned a function that maps variable assignments pairing George Bush with 'x' to *True*; and mapping variable assignments pairing Hilary Clinton with 'x' to *False*. Exploiting this parameter, we can characterize

[5] See in particular current work by Eklund (2005); Rayo (2004) on aspects of this relationship.

[6] Those interested in its development, and particularly on the current debate about its role within an account of vague language, should consult Williamson (1994) and Keefe (2000).

quantification: 'For all x, x is male' will be true at a variable assignment iff 'x is male' is true with respect to every variable assignment.

More sophisticated intensional operators can be introduced using accessibility relations on the class of possible worlds, or of variable assignments. Let a world w P-access u iff the laws of nature of w are not violated in u. Then we can set up the notion of *Physical Necessity*, PN, with the effect: 'PN, Britain is in the Arctic' is true at w iff 'Britain is in the Arctic' is true at u for all u that w P-accesses.

Encapsulating all this, a *model* for a language containing these two kinds of indices will take the form:

$$\langle D, W, V, R, F, a\rangle$$

where D is a set (understood as the *domain of all objects*) W is a set (understood as the *set of all possible worlds* and a a point within that set (understood as the *actual world*, V is a set (understood as the set of all variable assignments), R contains a string of *accessibility relations* and F is the *interpretation function* that assigns to expressions of the language appropriate *intensions*—that is, functions from pairs of elements drawn from W and V to extensions based on D.[7] (We shall use the notation $(m)_D$ to pick out the domain of m, $(m)_W$ to pick out the set of worlds in m, etc.) Call the space of all such models *semantic space*.

With this in place, we can define the unrelativized notion of TRUTH on the basis of the parameterized *truth under interpretation F at world w on variable assignment v*. First, we define 'S is true on model m'. The model will provide us with the 'designated interpretation' of the language and an 'actual world', but what of the variable assignments? It makes no obvious sense to think of picking out a 'privileged' variable assignment; so instead we de-parameterize by *generalizing*. The result:

> S is true on model m iff for all variable assignments v based on $(m)_D$, S is true under $(m)_F$ at $(m)_a$ and v.
>
> S is false on model m iff for all variable assignments v based on $(m)_D$, S is not true under $(m)_F$ at $(m)_a$ and v.

We get truth *simpliciter* by picking out a 'designated' model: the one where the $(m)_a$ is truly the actual world, where $(m)_F$ is the correct interpretation of the language, etc. The job of metasemantics is exactly to help us pick out this designated model. Given this, we can characterize truth and falsity *simpliciter* in the obvious way:

> S is TRUE iff S is true on m, where m is the designated model.
>
> S is FALSE iff S is false on m, where m is the designated model.

To get a model-theoretic characterization of logical consequence, we start with the intuitive idea that ϕ is a logical consequence of ϕ if *no matter what*, if Γ is true then so is ϕ. Actually, this won't quite do: consequence will be vacuous if we are allowed, say, to re-interpret 'and' as *or* (for then we can create counterinstances even to the move from 'A and B' to 'A'). Therefore, we pick out a subspace Λ of semantic space: *the logically admissible models*. These are those that agree with the designated model on the interpretation of a certain special range of expressions: for example, on the interpretation of *logical constants* such as 'and', 'or' and quantifiers such as 'exists' and 'for all'.

Thinking of consequence as truth-preservation under logically admissible models in this way, the following 'global' characterization of consequence is natural (Williamson, 1994, ch.5):

[7] This is a 'Carnapian' model in the sense of Chapter 5. See also the discussion in Appendix C.

> ϕ is a GLOBAL LOGICAL CONSEQUENCE of Γ iff For all models m within Λ, whenever each element of Γ is true on m we have ϕ true on m.

An alternative characterization works with the original parameterized semantic properties:

> ϕ is a LOCAL LOGICAL CONSEQUENCE of Γ iff For all models m within Λ, for all worlds w in $(m)_W$ and variable assignments v drawn from $(m)_D$, whenever each element of Γ is true under $(m)_F$ at w and v we have ϕ true under $(m)_F$ at w and v.[8]

On both these setups, we will have 'failures of bivalence' in a certain sense. In the actual world and on the intended interpretation, 'x is male' is true on some variable assignments and false on others. It is thus not TRUE, but it is not FALSE either.

Delineations

The above discussion illustrates general tactics which we can apply in many cases. The thought concerning indeterminacy is one implementation of these ideas. To motivate the extension, consider the following pair of operators:

> Definitely, a person with no hairs is bald.

> In some sense, Alfie is bald.

The two operators 'in some sense' and 'Definitely' have formal analogies to intensional operators such as 'necessarily', discussed above. Lewis suggests we treat them as such, introducing a new parameter—that of a 'delineation'—to accommodate them. In what follows, a delineation may be thought of as a 'way of drawing boundaries', where (one would think) the meaning-fixing facts do not fix such things.[9] 'Definitely S' will now be treated as requiring S to be truth at *all delineations*. Now this might make it look as if we only need delineations which are (intuitively) 'ways of making precise' our language as is stands: ones which draw the boundaries for predicates *compatibly with the meaning-fixing facts*. In fact, I think we need to revise the setup a little: I discuss this in Appendix A. For the moment, let's assume the contrary: the designated model will include a set of delineations Δ corresponding to all and only the ways of 'sharpening' the meanings of our language compatibly with the meaning-fixing facts.

Now, just as with variable assignments, it seems that we have no way to pick out a 'privileged' delineation. The natural tactic is to generalize, just as in the earlier case:

> S is true on m iff For all delineations d in $(m)_\Delta$ and for all variable assignments v drawn from $(m)_D$, S is true under $(m)_F$ relative to $(m)_a$, at d and v.

[8] As explained in Appendix A, global and local consequence as currently characterized come apart when we consider argument forms involving free variables.

[9] Formally, we can take delineations to be structureless points: the idea of them as a 'way of drawing boundaries' comes form their interaction with the interpretation function. Intuitively, the extensions assigned to a predicate P by f relative to a delineation d, will have a certain boundary; and if we substitute d' for d we get an extension with slightly different boundaries. I shall continue to use 'delineation' in a way that confuses delineations proper with the image of the delineations under the mapping given by the interpretation function.

Lewis (1970a) attributes to delineations much richer structure. This does not seem to be exploited in his formal treatments, and Keefe (2000) argues, to my mind convincingly, that there will be no principled extension of this kind of structure to vague predicates such as 'nice'.

TRUTH *simpliciter* is once more truth at the designated model.

Now consider an intuitively borderline case of a bald man—Alfie, say. By our understanding of what a delineation is, on the designated model there will be delineations relative to which he is bald, and delineations relative to which he is not. By our characterization of TRUTH, 'Alfie is bald' is not TRUE. Similarly, this is not FALSE. We have it that 'Alfie is bald' is not TRUE nor FALSE. Formally, this 'failure of bivalence' is exactly like the one we observed occurring in the case of the free-variable sentence 'x is male'. It has the same root cause: de-parameterizing via a generalization, rather than by picking out a privileged point.

Semantic properties

Let us review the unrelativized semantic properties that the supervaluational framework allows us to attach to words and sentences.

First, at the level of sentences, we have the notion of TRUTH that results from generalizing across all delineations: a property often called *supertruth*. Equally we have FALSITY, or *superfalsity*. We can also define INDETERMINACY as a third status: S will be INDETERMINATE iff it is neither TRUE nor FALSE. We have already seen two examples of sentences that arguably are INDETERMINATE in this sense: the free-variable formula 'x is male', and the case of borderline-bald Alfie: 'Alfie is bald'.

But the central focus of the inscrutability we are looking for is at the level of word-reference, rather than the truth-status of sentences. Here we can define three analogous statuses for other semantic notions, again in two steps:

> N refers to O on m iff For all delineations in $(m)_\Delta$, and all variable assignments v drawn from $(m)_D$, N refers to O under $(m)_F$ at $(m)_a$ and v and d.
>
> N fails to refer to O on m iff For all delineations in $(m)_\Delta$, and all variable assignments v drawn from $(m)_D$, it is not the case that N refers to O under $(m)_F$ at $(m)_a$ and v and d.

To get de-parameterized notions, we have:

> N REFERS TO O iff N refers to O on m where m is the designated model
>
> N FAILS TO REFER TO O iff N fails to refer to O on m where m is designated.

Again, there will be a third status: We say that N PARTIALLY REFERS TO O iff it neither REFERS nor FAILS TO REFER to O.[10]

Inscrutability within supervaluationism

Inscrutability arises when there is no choosing between a range of interpretations of a language. Taking these to each attribute classical extensions or intensions to expressions, we can encapsulate this whole range within a supervaluational setting by pairing each interpretation in the range with a *delineation*—the semantic value assigned to e according to the interpretation I is now assigned to e by the supervaluational semantics relative to the delineation d_I. Thus, where inscrutability arises, we have a canonical method for constructing a supervaluational semantics.[11]

Inscrutability of reference occurs when the range of 'intended' interpretations conflict over the semantic value assigned to a given expression. A singular term N, for example, might be assigned O by

[10] For the notion of partial reference, see Field (1974). Notice that, just as free variables allow us to construct INDETERMINATE sentences, variables themselves PARTIALLY REFER to every object.

[11] The idea of handling inscrutability within a supervaluational setting originates in Field (1973, 1974).

one interpretation, and O' by another. Within the overall theory, the term denotes O relative to one delineation, and O' relative to another. According to the definitions above, this is a case where a term t does not REFER at all. Rather, it PARTIALLY REFERS to a range of objects O and to O'.

The supervaluationist framework thus gives us a way of handling inscrutability, and relating it to semantic properties of expressions. In this way we can describe the effect of Quinean 'gavagai' arguments (to be discussed in Chapter 4): they try to show that 'Peter' PARTIALLY REFERS to a rabbit-stages, and to rabbit-parts, and to various other rabbit-related entities. Radical inscrutability of reference (to be discussed in Chapter 5) is the thesis that any singular term N PARTIALLY REFERS to every object whatsoever.[12]

[12] Some 'supervaluationist' settings involve machinery far richer than that described here, and will not naturally generalize to the case of inscrutability. The use of partial models in Fine (1975) and Kamp (1975) is an instance of this.

3.2 A theory-shadowing treatment of inscrutability

When we formulated interpretationism in terms of theory shadowing paraphrases in Chapter 2, we did not consider what would happen if more than one theory was 'selected'. Our implicit assumption was that we would find a single selected meaning-fixing theory; this encourages the supervaluationist treatment of inscrutability just described, for there we show how to construct a single overall meaning-fixing theory.

However, our original formulation identified meaning fixing theories with the best theory of the relevant data. If we simple-mindedly carry over this identification, we will end up with *multiple* selected theories in cases of inscrutability. The present section examines what happens if we allow this to happen.

The theory-shadowing paraphrase rendered '*p*' as 'According to best semantic theory, *p*', where this in turn was explicated as follows:

$$\ulcorner p \urcorner \rightarrowtail \ulcorner \forall x,y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash *p*)\urcorner$$

Notice that in order to receive a true paraphrase, a semantic statement must follow from each selected semantic theories. Inscrutability of reference occurs where the selected theories disagree among themselves. If *p* is a semantic claim over which the selected theories disagree, then we have the formalized versions of the following:

$\neg$According to best semantic theory p

$\neg$ According to best semantic theory $\neg p$.

That is, in such cases we will have an 'incompleteness' within a theory-shadowing account, analogous to those we have seen previously, within a quasi-fictionalism about set theory and within Rosen's fictionalist account of modality.[13]

If all optimal candidate semantic theories fall under 'selected', inscrutability is manifested in *incompleteness* within the theory-shadowing paraphrase. This suggests that we may be able to describe the phenomena using the kind of machinery we introduced to handle incompleteness puzzles quite generally.

To fix ideas, let us make up a word "Lonis", and suppose there is no choosing between candidate theories that assign *London* to "Lonis", and those that assign *Paris* to it. Let us suppose further that a pair of theories embedding the respective reference schemes are the only two 'successful' theories by interpretationist lights. Call the former θ_1 and the latter θ_2.

To get a grip on which of the paraphrased sentences hold and which do not, we can exploit the following equivalence:[14]

$$\theta_1 \vee \theta_2 \Vdash p \Longleftrightarrow (\theta_1 \Vdash p) \wedge (\theta_2 \Vdash p)$$

Quite generally then, *p* will be paraphrased to something true, iff it follows from disjunctive theory; and incompleteness within the former theory correspond with incompleteness within the disjunctive theory.

Some categorical statements about reference can follow from the theory. For example, if both theories agree that 'Madrid' REFERS TO Madrid, then clearly disjunctive syllogism will let us derive the categorical statement:

'Madrid' REFERS TO Madrid.

[13] An even closer analogy is with if-thenist accounts of mathematics, where the paraphrase is to a universalized conditional $\forall X(PA(X) \rightarrow p(X))$ if PA is not categorical (e.g. on a second order framework with Henkin semantics) then we find the same issues.

[14] The left-to-right direction holds since $\theta_1 \Vdash \theta_1 \vee \theta_2$, and $\Vdash$ is transitive. The right-to-left direction follows from disjunctive syllogism. Given $\theta_1 \vee \theta_2$, from each disjunct *p* follows, so *p* follows from the disjunction.

But the disjunctive theory does not, in general, deliver categorical statements about reference. Rather, we get disjunctive principles such as:

> 'Lonis' REFERS TO London $\vee$ 'Lonis' REFERS TO Paris

in this case, the lack of resolution impacts on the sentential level; we find:

> ('Lonis is in England' is TRUE iff London is in England)$\vee$
> ('Lonis is in England' is TRUE iff Paris is in England.)

Since one of the right-hand sides here is true and the other is false, we cannot conclude anything categorical about the truth-value of the sentence: the disjunctive theory (and hence our theory shadowing paraphrase) is *incomplete* at this point. Now consider a case where the right-hand sides of both biconditionals are true:

> ('Lonis is in Europe' is TRUE iff London is in Europe) $\vee$
> ('Lonis is in Europe' is TRUE iff Paris is in Europe)

Since both right-hand sides are true, disjunctive syllogism gives that, unconditionally, 'Lonis is in Europe' is TRUE.

Fact

Within supervaluational treatment the idea that there is 'no fact of the matter' about what a term refers to was explicated by attributing to the term a definite semantic relation to the objects in question: PARTIAL REFERENCE. The theory-shadowing account does no such thing: rather, cases where there is intuitively 'no fact of the matter' about reference correspond to incompleteness within the theory-shadowing paraphrase. This allows us to bring in the kind of machinery we have already found use for in describing incompleteness puzzles within the general fictionalist and quasi-fictionalist setting.

Recall the operator FACT which was introduced in §2.2. Recall that, when we have an incomplete fiction, we get sentences p such that

$$\neg\textsc{fact}[p] \vee \neg\textsc{fact}[\neg p]$$

(For example, in the case of a mathematical fictionalism, the generalized continuum hypothesis might be such an S). In the current context, we can introduce the relevant operator by the following biconditional:

$$\textsc{fact}[p] \Leftrightarrow \forall x,y(\textsc{selected}(x) \wedge \textsc{encyclopedia}(y) \wedge x \cup y \Vdash {}^*p^*)$$

In cases where the theory falls silent or multiple selected theories disagree (e.g. over whether 'Lonis is in England' is true) we will have

> $\neg$ FACT['Lonis is in England' is TRUE] $\wedge$
> $\neg$ FACT['Lonis is in England' is not TRUE]

Indeed, we can formulate this directly at the level of reference:

> $\neg$ FACT['Lonis' REFERS TO London] $\wedge$
> $\neg$ FACT[$\neg$ 'Lonis' REFERS TO London]

As noted previously, in the presence of incompleteness we cannot move from FACT[$p \vee q$] to FACT[p] $\vee$ FACT[q]. A striking illustration of this feature, in the case at hand, is that even though there's no fact of the matter over whether 'Lonis is in England' is TRUE or FALSE, and so we do *not* have:

FACT['Lonis is in England' is TRUE] $\vee$ FACT['Lonis is in England' is FALSE]

we still have the following:

FACT[('Lonis is in England' is TRUE) $\vee$ ('Lonis is in England' is FALSE)]

Indeed, given that the candidate semantic theories are classical, every instance of bivalence for closed sentences has a true paraphrase; and more generally, the theory shadowing account will conserve all the formal features of the semantic properties of the candidate semantic theories (of which bivalence is an instance, if the candidate theories are all classical). For if these properties hold with respect to any semantics of the candidate kind, then they hold in particular in each of the selected theories; whereby we can derive them by disjunctive syllogism from the overall theory.[15]

Formulating inscrutability of reference

We can use the 'FACT' operator to unpack our original characterization of inscrutability as a case where there 'is no fact of the matter' about what a given term refers to. One who follows Quine on 'gavagai' might maintain:

$\neg$ FACT["Peter" REFERS TO a temporal slice of a rabbit] $\wedge$
$\neg$ FACT[$\neg$ "Peter" REFERS TO a temporal slice of a rabbit]

To formulate the idea of *radical* inscrutability of reference, we really want to be able to *quantify into* FACT contexts. We want to say that there is no object O such that it is a fact that N fails to refer to it. That is, we want to formulate the claim as follows:

$$\neg\exists x\textsc{fact}[\neg N \text{ refers to } x]$$

However, we have as yet no machinery to understand what quantifying into such contexts amounts to. Because of one of the idiosyncracies of the setup outlined in §2.4, however, we can give an account of such a notion. In the context of a Lagadonian meaning-building language, the boundaries between substitutional and objectual quantification get blurry (as each object has a unique name—itself), and we exploit this in the following:

$$\textsc{fact}[\phi] \text{ holds of the objects } \bar{o}$$

$$\Leftrightarrow$$

$$\forall x,y(\textsc{selected}(x)\wedge\textsc{encyclopedia}(y)\wedge x\cup y \Vdash \ulcorner *\phi(\bar{o})*\urcorner)$$

That is, the open sentence FACT[$\phi(\bar{x})$] holds of a sequence of objects (assigned to the variables) iff, translating the matrix into Lagadonian, and then filling the places of the variables with those objects, we get something that follows from best theory.

We have seen two ways of handling inscrutability of reference within a broadly interpretationist setting. Supervaluationism allows us to construct a single overall meaning-fixing theory out of a range of optimal classical semantics. Or, simply letting each of the optimal theories be selected, we can allow

[15] McGee and McLaughlin (1994) describe a non-classical semantic treatment of a vague language that supports bivalence. However, their setting differs substantially from that in view here. They draw a sharp distinction between 'correspondence truth' for which bivalence fails; and 'disquotational truth' or 'pluth' for which bivalence holds. On the latter notion, they write that it is "a way of using the word 'true' whose fundamental governing postulate is the disquotation principle" (p. 217). By contrast, the treatment here need not support disquotation at all, since the metalanguage can in general be perfectly precise.

the theory-shadowing paraphrase to stand as it is. On this latter approach, inscrutability of reference is identified as arising from the kind of incompleteness familiar from other theory-shadowing accounts.[16]

[16] What happens if we formulate the theory-shadowing paraphrase in terms of an *existential* quantification over selected theories on the liberal understanding of 'selected'? The result will be the analogue of dual of supervaluationism, known as subvaluationism. Characteristic features are that some sentences will be true, and will also be false; but that no sentence is simultaneously true-and-false. See Hyde (1997) for details.

The universal treatment corresponds to a disjunctive theory, as we have seen; does a similar correspondence hold for the existential treatment? A *conjunctive theory* $\theta_1 \wedge \theta_2$ where the theories give competing assignments of truth-values will not do; for unlike the existential treatment, this represents certain sentences as both true and false. What is needed is a logic which fails to satisfy certain classically valid multi-premiss reasoning; in particular, $p, q \Vdash p \wedge q$. Given this, we can have $\{\theta_1, \ldots, \theta_n\}$ can be our theory. The failure of multi-premiss arguments is a well-known feature of the logic sustained by subvaluational semantics, and is paralleled in the supervaluational case by a failure to sustain classically valid patterns in a logic allowing multi-premiss conclusions.

3.3 Object-language expressions of inscrutability

We have seen that the supervaluational and incompleteness proposals for handling inscrutability allow formulation of inscrutability *from the perspective of the theorist*, using metalinguistic notions such as PARTIAL REFERENCE and FACT. Do they allow inscrutability to be stated from *within* a language that is itself inscrutable? If the inscrutability of a language could not be formulated within it, all sorts of tricky issues arise. Fortunately, they can be avoided, for both the above treatments allow us to characterize object-language operators that suffice to express inscrutability.

A famous feature of the supervaluationist setting is its ability to characterize an object-language operator DEF, which is the object-language equivalent of the supervaluationist's metalinguistic TRUTH (supertruth). The characterization is simple:

> 'DEF $[F\bar{x}]$' is true relative to variable assignment v and delineation d iff '$F\bar{x}$' is true relative to variable assignment v on all delineations.

The upshot is that inscrutability can now be expressed in purely object-language in terms. The radical inscrutability claim, that a name such as "Londres" PARTIALLY REFERS to everything, now finds expression in the claim that "Nothing is definitely not identical to London":

> $\neg\exists x$ DEF $[x \neq \text{Londres}]$

It will help to modify one aspect of the framework. (I give considerations in favour of the following in Appendix A). I recommend we replace the assumption that the optimal candidate interpretations are in one-to-one correspondence with the delineations within the intended model. Rather, I hold that we should have delineations giving 'extreme' and unintended extensions and intensions to expressions. We then add to our models an extra element: the set of *sharpenings*, which are the subset of the delineations that correspond to some one optimal candidate interpretation. The supervaluationist's definitions of TRUE should be modified accordingly: in characterizing what it is for a sentence to be true in a model, we now generalize, not over all delineations, but just over the sharpenings. We also need to modify the characterization of DEF. To this end we introduce a matching accessibility relation, S, which is (a) reflexive (b) such that every sharpening accesses every other. The characterization of 'Definitely' then becomes:

> 'DEF $[F\bar{x}]$' is true relative to variable assignment v and delineation d iff '$F\bar{x}$' is true relative to variable assignment v on all delineations which are S-accessed by d.

At first glance the theory-shadowing account only has a way of expressing the claim *in the theorist's metalanguage*, via claims such as $\neg\exists x$FACT$[\neg N$ REFERS TO $x]$. Even this relied on some idiosyncratic features of the setup—without a Lagadonian meaning-building language, we would have to rely on disjunctions and explicitly substitutional quantification to express the claim.

The aim here is to show that the theory-shadowing account can allow a language to express its own inscrutability. The first thing to do is to bring across much of the framework of supervaluational semantics, in the modified form given above. We have delineations corresponding to all sorts of crazy ways of assigning intensions and extensions. However, whereas supervaluationist models of the language contained a privileged set of delineations, the sharpenings, each optimal candidate theory will now pick out a single privileged delineation.

The claimed inscrutability would mean that we won't know which of the various semantic theories (i.e. a designated model picking out a single privileged delineation) is the right one to choose. Inscrutability will be a case where there are multiple 'best models', differing only over which delineation is actualized. Suppose that these are $S_1,\ldots,S_n$. Each comes with a privileged delineation it thinks is actualized; write these $\delta_1, \ldots, \delta_n$ respectively. Then we can *add* to each of these theories an accessibility relation M-accesses, characterized by:

> d M-accesses d' iff d and d' are each one of the δ_i's, or $d = d'$

Call the supplemented semantic theories $S_1^*, \ldots, S_n^*$. We can use the M-access relation to define a type of $\textsc{Def}_M$ operator which will enable us to express inscrutability:

> '$\textsc{Def}_M\ [F\bar{x}]$' is true relative to variable assignment v and delineation d iff '$F\bar{x}$' is true relative to variable assignment v on all delineations which are M-accessed by d.

Since the set of sharpenings on the supervaluationist account are exactly the set of δ_i above, the two relations S-access and M-access coincide exactly, and so allow the same expressive power. We can parallel exactly the moves which the supervaluational treatment makes using their DEF operator based on the S-accessibility relation. In particular, given radical inscrutability, it will be TRUE on each S_i^* that:

> $\neg\exists x\ \textsc{Def}_M\ [x \neq \text{Londres}]$

This shows us that we can *construct* a semantic theory with the resources to express its own inscrutability. It also allows us to give a general characterization of when a $\textsc{Def}_M$ operator is tracking inscrutability: when its defining accessibility relation coincides with the characterization of M-accessibility given above.[17]

[17] Notice that this (model-theoretic) approach seems inapplicable within a Davidsonian truth-theoretic setting. I do not see any straightforward way of introducing the 'Def' operator into the object-language in such a setting. (For example, we might look at defining a provability predicate for a (disjunctive) truth theory in order to formulate suitable axioms. For reasons given earlier, we need to add an 'encyclopedia' to the semantic theory to get the desired results. It is far from clear that the encyclopedia will be axiomatizable in the way needed.

3.4 Comparing the two accounts

How does the more idiosyncratic treatment of inscrutability suggested by the theory-shadowing analysis of semantic facts compare to the familiar supervaluational treatment of indeterminacy? Let us review their basic features.

Interpretationism delivers a range of "successful" semantic theories. Each semantic theory contains a certain assignment of extensions or intensions to expressions. We can either embed this information within a supervaluational treatment, or exploit the classical theories within the theory-shadowing paraphrase directly, treating inscrutability as a form of incompleteness.

In the former case, we assign statuses to the agent's sentences: as TRUE, FALSE, or INDETERMINATE; and relations that her words bear to objects: REFERENCE, NON-REFERENCE and PARTIAL REFERENCE. In the latter case, each classical semantic theory embeds a proposal about the application of the theorist's TRUE. As previously noted, we can get a handle on the theory-shadowing paraphrase by considering the theory that disjoins all the successful candidate theory. The disjunctive theory either entails '*S*' is TRUE; or entails '*S*' is FALSE; or remains silent. The silences correspond to incompleteness within the theory, and can be expressed, by the theorist, using the FACT operator. In general, the incompleteness account's

FACT[*S* is TRUE]

can be matched with the supervaluationist's

S is TRUE

Where the incompleteness account says there are no FACTS about TRUTH or FALSITY, the supervaluationist says that, on the contrary, the sentence is INDETERMINATE—neither TRUE nor FALSE.

The differences are marked. Given a classical backdrop, bivalence holds for the disjunctive-style theory but not the supervaluational one. Clearly, then, these two proposals are distinct, since they give rise to extensionally diverging overall analyses of semantic properties.

Given that the two kinds of approach are distinct, what are their relative merits? We first outline the areas of contention, and briefly assess the costs and benefits.

1. **Bivalence vs. failures of bivalence**
   We have noted that the theory-shadowing proposal is typically *conservative*. The incompleteness account's truth predicate retains all the formal features of the background theory: including (given a classical backdrop) all instances of bivalence. At the heart of the supervaluationist picture, on the other hand, is the thought that INDETERMINACY (/PARTIAL REFERENCE) is a state incompatible with TRUTH and FALSITY (/REFERENCE and FAILURE TO REFER).

2. **Logical revisionism?**

   Unlike the supervaluationist treatment, the incompleteness account of inscrutability is entirely classical in its assumptions. Since the semantic theory it selects between are classical, when we come to define consequence, we find no threat of departure from classical logic. Formally, we find the setup analogous to a standard classical modal logic.

   On the other hand, the supervaluationist treatment uses the non-classical explications via a generalization over a set of points—the sharpenings. It has been alleged (Williamson, 1994; Keefe, 2000) that, when combined with GLOBAL LOGICAL CONSEQUENCE, deviations from classical logic arise. A complex debate then arises over the acceptability or otherwise of such revisionism.

   In Appendix A, I argue that even if we interpret consequence within the supervaluationist setting in a global way, as do Williamson and Keefe, we do not get any departures from classical logic.

Moreover, even if we did get the kind of features that Williamson and Keefe allege, these may not count as a revision. The rules for which we find counterexamples already have counterexamples involving open sentences, given the global characterization of consequence. If this is right, there is nothing to choose between the two accounts on grounds of conserving classical consequence.

3. **Expressing Inscrutability in the object language.**

   The construction of the operator 'Definitely' is a distinctive feature of the supervaluationist setting—and allows us to formulate sentences that characterize the inscrutability of a language within that language, and without using any semantic vocabulary.

   We have just seen, however, that we can ape this construction within the theory-shadowing setting. Again we have no grounds for discriminating between the setups.

4. **Generality**
   Supervaluationism is flexible, in that it embodies a standard technique for generating a new kind of semantics out of a given range of many-valued logics.[18] Clearly, the incompleteness account also generalizes in this way. Further, the latter can be extended to cases where the supervaluational account is inappropriate. For example, Davidsonian truth theoretic semantics are often formulated in terms of a restricted syntactic consequence relation—canonical derivability. In the absence of a model theoretic correlate of this notion, the supervaluationist cannot extract 'valuation functions' from a given successful semantic theory. On the other hand, the incompleteness approach will extend to this case.[19]

5. **Unity**

   There is a major question whether everything that could be thought of as a case of 'inscrutability' can be captured by the supervaluationist approach. We shall shortly review a variety of cases of inscrutability: the incompleteness approach looks better placed to handle this than the supervaluationist.

6. **Interaction with vagueness** One common deployment of supervaluationism is in a supervaluational semantics for vague language. Our focus on *classical* semantic theories for natural language is at the moment justified only on methodological grounds. One might hold that *from the beginning* we should compare the relative merits of semantic theories tailored specifically for vague languages—and supervaluational theories (perhaps cashed-out as in Kamp (1975); Fine (1975)) would be a strong contender for providing that framework. If so, then inscrutability will correspond to multiple successful supervaluational semantic theories. Now, on the theory-shadowing approach, inscrutability and (semantic) vagueness are cleanly separated. Semantic vagueness is given a semantic treatment; whereas inscrutability is handled metasemantically.

   If, however, we adopt the supervaluational approach, we shall have to say something about how vagueness-related models and inscrutability related ones interact. One option is simply to expand the number of sharpenings to include both those 'admissible' due to inscrutability and those due to vagueness (that is, given theories θ_1, θ_2 with respective sets of sharpenings S_1, S_2, we let the sharpenings of the combined theory Θ be $S_1 \cup S_2$). For example, it might be that various chunks of land corresponding to ways of drawing precise boundaries around London or Paris will be

[18] Beall and van Fraassen (2003)

[19] We would need to take care, however, in how we handle the encyclopedia. If the encyclopedia was originally thought to be a set of principles which entailed all non-semantic facts about the world, we can simply replace this by its (classical) logical closure.

> assigned as referent to 'Lonis' on different sharpenings. It would be much more satisfying to find a formal setting in which such potentially damaging interactions did not arise, but I have no space to investigate this here.[20]
>
> (One reason for wanting such a nuanced technique is the risk of damaging interactions between inscrutability and the theoretical resources of a semantic account of vagueness. One potential area is in the kind of *de re* use of 'Definitely' that is commonly felt to have a significance in diagnosing the seductiveness of sorites-style reasoning.[21] But in the context of inscrutability of reference, the distinction will collapse.)[22]

We have not yet seen any decisive reason to prefer one account over the other. On grounds of the availability of a construction of 'Definitely', and of conserving classical consequence, they are equally matched. The incompleteness account has a claim to be conservative on other matters: for example, in preserving bivalence. However, once it is appreciated that 'failures of bivalence' are formally analogous to those that occur with respect to free-variable sentences such as 'x is male', it becomes unclear in what sense this constitutes a *revision*.[23]

Generality and Uniformity, on the other hand, potentially count in favour of the incompleteness account, as does the fact that the supervaluationist has the burden of showing that her treatment of inscrutability will not damagingly interact with her account of vagueness (of course, if she is, for example, an *epistemicist* about vagueness, she will feel no pressure here). The incompleteness account has the advantage of neutrality here.

A marginal case can be made, therefore, for favouring the purely theory-shadowing account as the better of the two options.

3.5 Concluding remarks

That some limited form of inscrutability infects our language should not come as a surprise. There are all sorts of features of semantic theories as they are usually presented which we would not wish to regard as reflecting, or taking a stand on, the underlying semantic facts. In these final remarks I sketch one case of this. If the possibility of this kind of innocuous inscrutability is admitted, then a key question for inscrutability arguments to come later in the thesis will be: can the resulting inscrutability be regarded as of the innocuous kind? Or is something more worrying being claimed?

My example of innocuous inscrutability of reference focuses on the reference of predicates. Classical semantic theories assign sets of n-tuples as the referent of an n-adic predicates. For example, 'is father of' might be assigned the set of pairs:

$$\{\langle \text{Ozzy, Kelly}\rangle, \langle \text{Charles, William}\rangle, \langle \text{George, George W.}\rangle \ldots\}$$

A familiar fact about the set-theoretic representation of ordered pairs is that they can be constructed in a variety of ways. The most usual set-theoretic reduction of ordered pairs to unordered pairs is from

[20] Elia Zardini has suggested to me that a direction of progress would be to exploit the machinery that has been developed to handle higher-order vagueness (for present purposes, thought of as indeterminacy over what is determinate) within supervaluational theories of vagueness: in particular, that discussed in Fine (1975).

[21] For discussion of such distinctions, and their theoretical significance both in general and in the context of supervaluational theories of vagueness, see Fine (1975); Keefe (2000); Greenough (2003); Edgington (1997); Weatherson (2002). See also the brief discussion on page 120, below.

[22] Though, it is in fact quite tricky to maintain the distinction within a standard supervaluational framework. In Williams (2006a) I use this to argue for the "many" treatment sketched in Lewis (1993). The troubles may suggest that we find a substitute for the theoretical role being played by 'Definitely'. This is an area I hope to explore in future work.

[23] It is a nice question how we should think of arguments purporting to establish a contradiction from the denial of bivalence in application to the free-variable case.

Kuratowski (1921)

$$\langle a,b\rangle := \{\{a\},\{a,b\}\}$$

Clearly, there is nothing 'uniquely special' about this particular reduction. Equally good would be the following

$$\langle a,b\rangle := \{\{b\},\{a,b\}\}$$

or the version suggested by Wiener (1914):

$$\langle a,b\rangle := \{\{a,\emptyset\}\{a,b\}\}$$

If, as is often supposed, the semantic values of expressions are set-theoretic constructs, then there seems an irreducible element of arbitrariness in exactly which semantic values we pick.[24] Thinking that one of these constructions is a 'reference magnet' is surely too much of a bullet to bite; more attractive is to say that there is *no fact of the matter* in such cases.[25] This certainly fits our intuitive characterization of the inscrutability of reference—and the incompleteness account, at least, has the resources to depict it as such.[26] This kind of inscrutability of the semantic framework is, I claim innocuous: it would be worrying if it *did not* hold.

In addition to framework inscrutability of this kind, there are arguably various illuminating kinds of inscrutability. In these cases the inscrutability, rather than being disturbing, helps one to understand otherwise puzzling phenomenon. One possible example is that of Newton's concept of 'mass'. Field (1973) argues that detecting inscrutability in the term 'mass' (between the candidate referents *rest mass* and *relativistic mass*) can help us to rebut the charge that consideration of scientific revolutions render a referential semantic theory implausible. Unger's problem of the many, and vague language in general, may be cases where inscrutability of reference illuminates puzzling features of language.[27]

[24] Compare Benacerraf (1965). One way of responding to this puzzle, paralleling Wetzel's (1989) response to Benacerraf, is to posit *sui generis* semantic values—abundant *relations* for each predicate. We would then say that the arbitrariness issue identified here is a problem of set theoretic *representation* only. Some concerns: (1) postulating *sui generis* ontology is generally unattractive; (2) Unlike the mathematical case, we need not only the ontology, but also the additional *ideology* of an 'instantiation' primitive—in fact, we either need infinitely many such primitives, corresponding to each adicity of predicate, or else a multigrade primitive. (3) We would then have to face *metaphysical* concerns that the non-symmetric primitive of instantiation required is incoherent. See Dorr (2004).

In the general case, an n-tuple can be identified with a function from the first n natural numbers to extensions, so that

$$\langle a_1,\ldots,a_n\rangle := f : i \mapsto a_i$$

The value of a n-adic relation will then be a set of such functions. Notice that in the standard set-theoretic case, the functions themselves will be sets of ordered pairs, so the arbitrariness in pair-construction will still be present. In a theory taking functions as primitive, this may be finessed; but alternative representations may still be constructed, for example, by permuting the order in which the terms are given. Note that postulating *sui generis* functions will be susceptible to some of the criticisms given above: *functional application* would be a non-symmetric primitive notion.

[25] Cf. Hodes (1984) on reference magnetism. Williamson (1994) notoriously defends the idea that there can be facts about which precise extensions vague terms pick out. This to most people seems wildly implausible. To adopt such a view in the case at hand seems another level of implausibility beyond even Williamson's position—for it seems that there is *in principle* no way for one construction rather than another to be favoured. I think, therefore, that even epistemicists should be interested in an analysis of inscrutability of the kind I will offer. In discussion, Williamson has said that, if convinced by the arguments for radical inscrutability, he would extend his epistemicism to such cases also. He may therefore be willing to bite the bullet even on the 'innocuous' kinds of inscrutability just mentioned.

[26] This is one point at which the incompleteness account may gain in generality by comparison with the supervaluational treatment. For clearly the theory-shadowing treatment can regard theories exploiting different constructions of semantic values for relations as rival semantic theories, and presumably the set of optimal semantic theories will be closed under at least some of the ways of varying this. On the other hand, one would have to work hard to get this into the supervaluational framework, for the overall model will have to provide once-and-for-all criteria of admissibility for interpretation functions, or recipes for extracting truth-on-f from an interpretation function f which assigns semantic values to the atomic parts of language.

[27] See Unger (1980) and Weatherson (2004). McGee (2005a) explicitly presents the problem of the many as a form of inscrutability.

These two paradigms—innocuous inscrutability, and illuminating inscrutability—may not exhaust the kinds of inscrutability of reference we find. Our focus in the next few chapters will be on arguments for surprising instances of inscrutability, that neither have the innocuous feel of the former kind, nor yet give the impression of illumination. Indeed, they have been found paradoxical by many.

Chapter 4

Gavagai again

This chapter presents arguments that interpretationism (at least as currently formulated) is committed to what I call *division inscrutability*: inscrutability concerning what kind of entities general terms such as 'Rabbit' divide their reference over.

My discussion falls into two parts. In the first part, I discuss the form of arguments for division-inscrutability, some of the general metaphysical and methodological issues involved, and some important considerations that surface in commentaries on Quine's "argument from below". In the second part, I turn to a detailed presentation of three semantic theories corresponding to different ways of dividing reference, drawn from the literature on persistence in the philosophy of time, and analogous systems for analyzing extension in space. I then address objections that have been made in the debate on persistence that are relevant to our setting; and discuss how the objections made to Quine can be handled within the framework here developed.

Though our discussion is focused on potential inscrutability, much of the material here could be presented as engaging in the literature on persistence. Our considerations naturally raise the meta-philosophical question of what the subject-matter the debate on persistence (in particular, the worm theory/stage theory debate) is supposed to be: is it *purely* a matter of the appropriate semantics for English, as is sometimes implied?[1] If not, what kind of debate is it? If it is a debate about semantics, should not the kind of indeterminacy or inscrutability thesis here advocated be seriously considered by interested parties?

4.1 Quine on 'gavagai'

The second chapter of Quine's (1960) "Word and Object" contains his famous "argument from below". Quine presents his ideas in the context of *radical translation*. He considers the challenge to translate from scratch an unknown tongue: in particular, to translate the expression 'gavagai' that native speakers use. The scenario he describes is one where the expression is used (as we might say) to register the presence of rabbits—the kind of usage we find when a child exclaims *Rabbit!* as a rabbit runs past.

If one focused exclusively on such 'occasion sentences', presumably all one could do to elucidate their representational properties is to describe the states of affairs in which they are appropriately asserted (the states of the world which make the sentences 'true'). However, to describe the representational function of 'gavagai' as a general term, we need something more. We need something analogous to the assignment of an extension to 'is a rabbit': {Flopsy, Mopsy, Cottontail, Peter...}. It is here that Quine detects indeterminacy:

[1] Sider (1996a).

> Who knows but what the objects to which this term ['gavagai'] applies are not rabbits after all, but mere stages, or brief temporal segments, of rabbits? In each case the situations that prompt assent to 'gavagai' would be the same as for 'rabbit'. Or perhaps the objects to which 'gavagai' applies are all and sundry undetached parts of rabbits, again the same stimulus meaning would register no difference…
>
> A further alternative likewise compatible with the same old stimulus meaning is to take 'gavagai' as a singular term naming the fusion, in Goodman's sense, of all rabbits: that single though discontinuous portion of the spatiotemporal world that consists of rabbits. …And a still further alternative in the case of 'gavagai' is to take it as a singular term naming a recurring universal, rabbithood.
>
> (Quine, 1960, pp.51-52).

Quine argues that these kinds of translation-hypotheses are equally adept at explaining the subjects' speech behaviour. For him, this leads to the 'indeterminacy of translation'. I follow Field (1974) and Evans (1975) in reformulating the Quinean argument within the framework of semantic interpretation, rather than that of translation. In this setting, we present the Quinean case as follows. First, one can interpret 'gavagai' in one of the ways above: either as a predicate dividing its reference over rabbits, or over rabbit stages, or over undetached parts of rabbits; or as a singular term picking out the fusion of rabbits, or the universal Rabbithood. Second, that at the level of sentences (and thus in the predicated pattern of assent and dissent to sentences) there is nothing to choose between these rival interpretations.

When asking about the proper interpretation of a predicate, there are two distinct types of question one might ask. The first considers the 'logical form' of predication, or its proper *semantic analysis*. For example, one might think that in general one should regard a predication '*A* is *F*' as being underlain by one of the following:

- '*x* is *F*' is satisfied by the referent of '*A*'
- the referent of '*A*' is a member of the set which is the extension of '*F*'.
- the referent of '*A*' instantiates the property expressed by '*F*'

In each case, one finds a different idea about how predication should be framed: in the first case the "satisfaction" relation is taken as primitive. In the second case, predicates designate sets, and set-membership is taken as primitive. In the third case, properties are assigned to predicates, and property instantiation is taken as primitive. The general question, then, concerns the ontological and ideological commitments that surface in an analysis of predication, and the project is to determine which are to be built into the framework of a semantic theory. The semantic theories that we mentioned earlier take different stances on this: the Davidsonian treatment took satisfaction as primitive, whereas the model-theoretic semantics is framed set-theoretically.[2]

The second question to ask about predication is not a general 'framework' question at all: it is a question about to which objects a particular predicate applies. If we say that a predicate 'divides' its reference over those objects it applies to, we can ask, for example, what kind of objects 'Rabbit' or 'Gavagai' divide their reference over. Note that however one analyzes predicate-application, the question of reference-division can be posed: we can ask respectively what kind of objects satisfy '*x* is a rabbit'; what kind of objects are members of the set that is the semantic value of 'rabbit'; what kind of objects are instantiated by the property designated by 'rabbit'.

[2] There are alternatives: see Larson and Segal (1995, passim) for a number of formulations of a Davidsonian theory, for example, in terms of a property/instantion framework. The model theoretic framework too, could be implemented in a variety of ways: for example, in terms of a theory of functions rather than of sets.

Seen in this light, the ‘argument from below’ if successful would establish two different inscrutability results: inscrutability of analysis of predication, and division-inscrutability. The former says that there is no fact of the matter whether falling under ‘gavagai’ consists in being a member of the (set-theoretic) extension of ‘gavagai’; or rather instantiating the universal *Rabbithood* denoted by ‘gavagai’.[3] I think that we should include the ‘rabbit-fusion’ proposal within this class: the corresponding analysis of ‘A is F’ would be: *the referent of A is part of the referent of ‘F’*.[4]

Inscrutability of *reference division*, on the other hand, maintains that there is no fact of the matter whether the objects falling under the predicate (however that is to be understood) are rabbits, or instantaneous stages of rabbits, or undetached parts of rabbits.

4.2 Metaphysics and methodology

The ontological and ideological commitments of a theory may depend to a considerable extent on which of the various options just mentioned is sustained.[5] On the logical-form side, we find commitment variously to: properties and instantiation; sets and set-membership; primitive satisfaction; mereological fusions and the part-whole relation; and so forth. On the reference-division side, we find commitment to persisting objects, or merely to instantaneous stages, or even just to mereological atoms.[6] The former are putative commitments of predication; the latter are putative ‘Quinean’ commitments: what kinds of object lie within the range of the first order variables.

The ontological commitments of our theories thus depend in part on which, if any, of the various hypotheses hold. The claimed Quinean inscrutability will lead to it being inscrutable what the commitments of our theories are. Since the various alternatives have *different* commitments, it is hard to see how we could establish inscrutability without an assurance that the world contains the properties, sets, and various types of object called on by the respective theories. If one way of glossing predication, for example, has it that a sentence is true iff the object A instantiates a particular property P, and in reality there are no properties, then the upshot would be that the sentence is false. A more nominalistic gloss on predication might render the sentence true. Whether such a situation arises is clearly relevant to judging the merits of the two semantic proposals.

Some accounts of the relationship between linguistic analysis and metaphysics might blur this natural thought. But it is only the most radical theses of constitutive dependence of ontology on ‘conceptual scheme’ or the like, that would deny the possibility of the kind of scenario outlined above. I will assume in what follows that we can bring independently grounded metaphysics to bear on the current debate.

Considerations from *property-ontology* can be brought to the alleged inscrutability of logical form. If, for example, one rejects the existence of abstracta such as sets, viewing predication as committed to sets looks unattractive. At the least, it will commit one to a universal error-theory, since there will

[3] One can look at the literature on property-ontology to find more alternatives: for example, a theorist who viewed objects as bundles of tropes might analyze ‘x is a rabbit’ as *A rabbit-trope is part of A*. See Armstrong (1989) for discussion of these issues.

[4] The natural objection to this is that any *parts* of an object that falls under F will now also fall under F. What I take this to show is that in the case we cannot be neutral over issues over how reference is divided: in particular, we have to combine this framework with the *undetached parts* proposal, whereby ‘Gavagai’ applies to every undetached part of a rabbit.

[5] I will not be following the Quinean understanding of these notions, since I will be taking predication to incur ontological commitments to whatever kind of thing is involved in a proper analysis of predication, be that sets or universals. This is something urged by Armstrong (1978a,b), but is resisted by Quineans. See Rayo and Yablo (2001). (Notice that one way of resisting any ontological commitments from predication, compatibly with the Armstrongian view taken here, is to adopt a framework whereby ‘satisfaction’ or ‘application’ as a primitive. More radically, one could take the distribution of particular predicates as primitive: one could analyze ‘A is F’ as *the referent of ‘A’ is F*: there would be no general schema available. This corresponds to the view of the ‘Ostrich nominalists’.)

[6] If we read ‘atomic undetached rabbit parts’ for ‘undetached rabbit parts’ we would have the latter scenario. This is the form in which I will be discussing it.

be no set picked out by '*F*' of which the referent of '*A*' can be a member. Likewise, one might reject 'abundant' universals (e.g. universals other than the framework notions of microphysics and perhaps other physical sciences) that would be needed to make work the *general* analysis of predication in terms of Universals (cf. Armstrong (1978b)). One's general metaphysical view is unlikely to contain overmany candidate-types for the *general* analysis of predication.[7] Because of this, I will not focus in what follows on potential Quinean inscrutability of logical-form.

Turning to alleged *division* inscrutability, issues in *particular-ontology* may constrain our arguments. To have a chance of establishing Quinean inscrutability, we would need to find a metaphysical picture which included rabbits, rabbit-stages, and undetached rabbit parts. A preliminary question concerns the nature of *rabbits*, and how these entities relate to the other categories. Elsewhere in his writings, Quine describes his "conception of a material object":

> ...the material content of any portion of space-time, however scattered and discontinuous. Equivalently: any sum or aggregate of point-events. The world's water is for me a physical object, comprising all the molecules of H_2O anywhere ever. There is a physical object part of which is a momentary stage of a silver dollar now in my pocket and the rest of which is a temporal segment of the Eiffel Tower through its third decade.... Among the myriad ways, mostly uninteresting, of stacking up momentary objects to make time-extended objects, there is one popular favourite: the corporeal. Momentary objects are declared to be stages of the same body by considerations of continuity of displacement, continuity of deformation, continuity of chemical change.
>
> (Quine, 1976b, pp.124-125)[8]

I think this makes it plain what Quine would take to be *rabbits* to be: space-time worms that are *fusions* of all the instantaneous rabbit stages of a single rabbit-life-history, and equally of all the undetached parts caught up in that life-history. I shall call such entities *rabbit worms*.

The interesting thing about the picture Quine here sketches is that it seems to deliver all the various candidates: Quinean 'rabbits' (i.e. rabbit-worms), rabbit-stages, and undetached rabbit parts (URPs). Such a view is taken seriously in contemporary debate: for example, in the "four-dimensional" view of particular-ontology that Sider (2001), for one, describes and advocates. According to this view, the past and future are as real as the present (i.e. it stands opposed to *presentism*), and the layout of reality contains no 'tensed' features. Furthermore, for any plurality of objects there is a 'fusion' that has those objects as parts (*unrestricted fusion*). For any space-time region 'filled' with an object, there is a part of that object existing exactly at that space-time region (*the doctrine of arbitrary undetached parts*). The notions of fusions, parts and wholes within this world-view are characterized by a classical extensional mereology.[9]

Within this picture, the rabbit-worms that Quine thinks of as the natural candidates to be rabbits, are to be found. Since such entities are four-dimensional objects, they fill spatial regions at many times. By the doctrine of undetached parts, there are objects ('temporal slices') that exist exactly at those times and nowhere else. Equally, the worms and slices fill smaller spatial regions: indeed, if space is pointy they 'fill' points. By appeal to the same principle, we have objects existing exactly at those regions and parts: undetached rabbit parts. (We could of course work this the other way: starting with the mereological

[7] An interesting *residual* question has already been raised concerning the ontological and ideological commitments of other ways of cashing out non-symmetric relational predication: see Dorr (2004) for some of the issues arising, and compare p.62 above for brief discussion.

[8] Page references are to the version collected in Quine (1981).

[9] For the mereological notions see Simons (1987); for formulations of four-dimensionalism see Sider (2001). For further discussion concentrating on the interaction of mereology and location see Parsons (2005b).

atoms existing at a single space-time point, we can appeal to unrestricted fusion to ‘build up’ rabbit stages and rabbit worms).

Clearly the four-dimensional setting is extremely controversial, and many reject some or all of its theses. Some reject the framework assumptions: its eternalist and de-temporalizing presuppositions. Some accept the framework but reject either unrestricted fusion or arbitrary undetached parts. Those who deny the former principle include both moderates and radicals. The former accept that fusion occurs in the case of rabbits, chairs and planets, for example, but not in the case of arbitrary scattered collections of objects. Radicals—for example, the various near-nihilist and total-nihilist positions of van Inwagen (1990) and Dorr (2002) respectively—deny that there are any non-living composite objects, or any composite objects at all. Those who deny the latter principle (of ‘arbitrary undetached parts’) include ‘endurantists’ and believers in spatially extended simples.[10]

Particular stances here can undoubtedly provide a metaphysical resolution to the alleged division inscrutability. For nihilists, only certain (mereologically simple) undetached rabbit parts exist: *prima facie*, anything but the final Quinean proposal would lead to a universal error-theory.[11] And one can imagine other metaphysical positions that would select among the various options we consider. Clearly a review of how the Quinean programme works out with respect to all these different metaphysical settings is beyond the scope of our present discussion.

What we have seen is that there is a world-view—that of the Siderian/Lewisian four-dimensionalist—that encompasses all the rabbit worms, stages and undetached parts one could wish for. I shall presuppose it in what follows for three reasons: (1) I am personally inclined to look favourably on this picture; (2) it seems to be Quine’s own favoured view; (3) it allows us to develop the options and give the Quinean division-inscrutability argument its best chance of success. We should not forget, however, that an independently motivated metaphysics that undermine aspects of the 4D setting, might yet impact on division-inscrutability.

[10] See Parsons (2000) on the compatibility of endurantism with the four-dimensionalist’s framework assumptions.

[11] van Inwagen (1990) and Dorr and Rosen (2002) recommend an analysis of ordinary talk as involving plural reference to simples. The analysis faces some worries (for example, in how it will treat ordinary plural talk!) and so nihilists should be interested in the Quinean alterative to be constructed below.

4.3 Quine's arguments for division inscrutability

Quine's argument

Quine's methodology is to challenge his opponent to point to some piece of linguistic behaviour that discriminates between the translation-schemes he describes. To begin with, we have the use of the general term 'Gavagai!' as an occasion sentence: used to register the presence of a rabbit. In our own case, we might take such rabbit-registering uses of 'Rabbit!' as short for 'There's a rabbit!'. On each of the proposals, we can find such glosses. For the three that concern us here, these are:

Candidate	**rabbit-registering interpretation**
Rabbit worm	there is a rabbit-worm partially present
Rabbit stage	there is a rabbit-stage present
URPs	there is an undetached part of a rabbit present

As Quine is well aware, there is a more general challenge: to explain linguistic behaviour where 'Gavagai' is embedded in more complex contexts. For example, we have synchronic and diachronic identity statements: it is not immediately obvious how the stage and URP proposals can deal with the analogues of 'this is the same *F* again'. For it is not the case that we find the same stage at two temporally separate occasions: the rabbit-stage in the hutch today is distinct from the one that was there yesterday. In the URP case, even synchronic identity poses problems: if (intuitively) a rabbit is partially hidden behind a post, then we can point to its nose and tail and truly say 'this is the same Rabbit as that', but we are pointing to distinct URPs.

Quine notes that to put these forward as objections to the alternative reference-divisions, a particular interpretation/translation of object-language talk of 'sameness' is being presupposed.

> If by analytical hypothesis we take 'are the same' as translation of some construction in the jungle language, we may proceed on that basis to question our informant about sameness of gavagais from occasion to occasion and so conclude gavagais are rabbits and not stages. But if instead we take 'are stages of the same animal' as translation of the jungle construction, we will conclude from the same subsequent questioning of our informant that gavagais are rabbit stages. Both analytical hypotheses may be presumed possible.
>
> Quine (1960, p.72)

Quine's general point here, mapped into our setting, is that if we vary our interpretations of the 'apparatus of individuation'—in particular, identity symbols—then we can finesse apparent counterexamples to our alternative schemes of divided reference. The general point is good: twists in one place can be 'cancelled out' by compensating twists elsewhere.

Clearly, though, the challenge is more general: the Quinean inscrutability argument requires that we have no good grounds for discriminating between the options, no matter which sentence we choose. Quine never attempts a general result along these lines.[12] This incompleteness in the Quinean case is exploited by the two authors we now discuss.

Internal problems for the argument for division inscrutability

Gareth Evans' paper 'Identity and Predication' (1975) includes a subtle examination of Quine's arguments. Evans makes several different kinds of points against the inscrutability arguments Quine offers.

[12] Field (1974), discussing how to accommodate Quinean division inscrutability within a systematic semantic theory, also concentrates on identity statements (his purpose, however, is not to defend division inscrutability, but to show that *even granted division inscrutability*, a systematic semantics is possible).

Some, for example, involve clever *ad hominem* attacks on Quine.[13] Others involve disputing the Quinean methodology for detecting predication and divided reference in a language.[14] But the ones of most interest to us simply point to sentences that it is not clear that the Quinean can handle. We shall concentrate on Evans' discussion of the Quinean case for division inscrutability.[15] As Evans concedes in the final paragraph of the paper, he does not offer an assurance that *no* semantic proposal will vindicate Quine. His methodology is rather to 'try out' some plausible candidates and show that they break down. Since I will be developing a detailed proposal in an attempt to vindicate Quine (or at least put his case in its strongest form) it would be pointless to examine in detail the views that Evans tries out. What we can hope to usefully extract from the considerations of Evans and others is a sense of the kind of problems that a semantic theory must meet, and a range of test cases for a potential analysis.

The first set of Evansian objections concern the proposal whereby 'Rabbit' divides its reference over all and only undetached rabbit parts. Evans asks about the objects that supposedly fall under adjectives such as 'white'. He offers a trilemma:

- If all and only white things fall under 'white', then a brown rabbit with a white foot will satisfy "white rabbit".
- If all and only things which are parts of a white rabbit fall under 'white', then the conditions of "white rabbit" will be fine, but the truth-conditions of "white house" will be wrong.
- If all and only things which are parts of a white *thing* fall under 'white', we get white rabbits and houses: but we overgenerate. A white rabbit foot is part of a white thing: on the current proposal, we would have 'white rabbit' satisfied in the presence of a brown rabbit with a white foot.

The Evansian question for any attempt to vindicate Quinean division-inscrutability is the following:

(A) Show how the semantics allegedly supporting the division of 'rabbit' over URPs can cope with the interaction of adjectives and general terms.

Fodor (1993), inspired by Evans, provides another challenge for the URP proposal. Rather than looking to adjectives, he focuses on cases where we have predicates both for a thing and for some distinctive part of that thing. The case he chooses is a language containing both 'rabbit' and 'ear'. There is nothing that is both an ear and a rabbit, but any undetached part of a rabbit's ear (say, the ear itself) will be an undetached part both of a rabbit, and of an ear. Therefore, on a systematic Quinean proposal, it looks as if the following will be true:

> there is something that is both a rabbit and an ear.[16]

[13] See the discussion of the identity of indiscernibles at Evans (1975, p.113?) (page references are to the version in McDowell (1985)).

[14] The basic contention seems to be (1) that Quine misses the constraint of *simplicity* on semantic theory. (2) That predicates should divide their reference over objects to whose spatial and temporal boundaries we are sensitive. Evans seems to regard (1) as the more fundamental principle, and (2) as resulting from an application of this principle. I do not see how the connection is to be established. We discuss simplicity constraints in Chapter 8.

[15] Evans' arguments against differing analyses of predication seem to me unconvincing: he assumes the Quinean interpretation of 'White Rabbit' would be *White instance of Rabbithood*, whereas I would render it *Instance of Rabbithood and of Whiteness* (this is plausible at least for adjectives that are in the category Kamp (1975) calls 'predicative'). I take it that Evans' considerations have no force against this proposal. His objection to the rabbit-fusion proposal is that I mentioned above: that any part of a rabbit will count as a rabbit. As flagged earlier, I think this shows that such a proposed analysis of predication will stand or fall with the claim that it is permissible to take 'rabbit' to divide its reference over URPs.

[16] Fodor's concern is actually with the conjunctions "*A* is an ear and *A* is a rabbit" and "*A* is an ear and a rabbit". I think the existential generalization presents the harder case for the division inscrutabilist.

The strength of Fodor's objection is that there seems no way that we can assign to 'ear' an extension that does not include some rabbit parts: but any such overlap will render true the strange generalization given above. Thus, we have a second challenge:

(B) Show how the semantics allegedly supporting alternative divisions does not generate false positives in the context of a pair of general terms, one of which (intuitively) applies to a part of the other.

In considering the rabbit-stage proposal, Evans focuses on tense. He maintains that *either* one will end up with a situation where if there is any stage of the rabbit which satisfies 'F', any stage of the rabbit at any time will satisfy 'F'; *or* one will be able to render true 'a rabbit was running' where the rabbit no longer exists. Here I am convinced by the rebuttals of Evans' specific arguments in Wright (1997) and Richard (1997). The general challenge is well taken however: a systematic development of the stage proposal will have to show how it is compatible with tensed ascriptions.

(C) Show how the semantics allegedly supporting alternative divisions can cope with tensed predications.

Having noted these concerns, I think that the best tactic to meet these concerns is to develop a systematic semantic proposal for each candidate 'division' of reference. We can then evaluate whether the challenges (A-C) that we have extracted from Evans and Fodor cause problems. This is the burden of the second part of this chapter.

4.4 Worms, stages, and undetached parts

I now spell out three candidate semantic theories for (a fragment of) English. In the first, 'rabbit' divides its reference over four-dimensional rabbit-worms; in the second, it divides its reference over three-dimensional (instantaneous) rabbit-stages. In the third, it divides its reference over one-dimensional undetached rabbit parts or 'dots'.[17]

The task is made infinitely easier by the development, in the literature on theories of persistence, of semantic theories that suit our purposes exactly. In particular, I will appeal to the basic semantic frameworks of the perdurance theories of persistence—"worm theory", and the semantic framework of what Haslanger (2003) calls the *exdurance* theory of persistence—the "stage theory" advocated by Sider (1996a, 2001) and Hawley (2001). For terminological convenience, I use 'perdurance' and 'exdurance' for the *metaphysical* theories of what it takes for an ordinary object to persist through time; I will use 'worm theory' and 'stage theory' to refer to associated *semantic* views.

Some inscrutability arguments are based on formal results proving that candidate semantic theories are 'sententially equivalent'—assign the same semantic values to sentences. I do not believe this to be possible here.[18] My methodology, rather, is one of providing a semantics for a *fragment* of a language: albeit a rather rich one. In each case, it will be a first-order intensional language with time and world indices, allowing modal and temporal operators to be defined. A full defense would require extensions to a setting adequate for natural languages as a whole, such as is arguably provided by the 'general semantics' discussed in the next chapter. However, the present fragments provide more than enough material for the key issues over division inscrutability—including those due to Evans and Fodor—to be formulated.

Worm theory

The perdurance theorists of persistence have two central theses—first, that ordinary objects like candles and rabbits are *perduring* space-time worms; secondly, such things undergo alteration in virtue of their stages successively possessing different properties. Haslanger outlines the view:

> On the perdurantist's conception of persistence, an object persists through time in a way analogous to how an object is extended through space. The candle is spatially extended through its 7-inch length . . . by having parts at the different regions. Likewise, according to the perdurantist, the candle is extended through time. . . by having parts or stages at different times. . . . The notion of perdurance provides the resources for a relatively straightforward account of alteration: . . . the persisting candle is composed of temporal parts or stages that only briefly exist; distinct *candle-stages* are the proper subjects of incompatible properties, *being straight* and *being bent*, and the temporal composite which consists of the stages is the subject of persistence. . . . On this account, persisting things are temporally extended composites, also known as . . . space-time worms
>
> (Haslanger, 2003, p.318)

A semantic theory that accords with the perdurantist theory of persistence, then, would be one where 'rabbit' divided its reference over rabbit-worms. Obtaining such a semantic theory is not without its

[17] I do not think that there would be any objection to extending the account respectively to non-atomic rabbit parts, or to more-than-instantaneous stages, but I shall not discuss this here.

[18] Wolfgang Schwarz (2005a) makes this claim for versions of the stage and dot theories below; but his versions have richer semantic structure, and it is not clear that they can be described as involving predicates whose reference is divided over undetached rabbit parts/stages rather than *sets of* such parts/stages. See also the discussion archived at Schwarz (2005b).

difficulties. In particular, it looks as if we have to discern two forms of predication, depending on whether the predicate at hand applies to a thing throughout its existence (arguably including: rabbit, chair, person); or can apply to a thing at some times and not at others (paradigmatic examples include 'runs', 'is green', 'wobbles', etc).

Following Parsons (2005a), let us divide predicates into two sorts: s-predicates (e.g. 'is a rabbit', 'is a chair', 'is a person', etc) and c-predicates (e.g. 'is running', 'is green' etc).[19] Parsons presents his theory taking the notion of s- and c-predicates as primitive, and deals only with the monadic case. To generalize to the polyadic case, we shall take predicates to be given with an assignment of sorts (s and c) to their name-places. An s-predicate is then a monadic predicate whose name-place is of sort s, and a c-predicate one whose name-place is of sort c; but more generally, an n-place relation might have k places of sort s, and $n-k$ places of sort c.

I will present the worm theory *together with a counterpart theoretic treatment of modal operators*. There is no reason why the world-shifting operators should not be treated in other ways, but I include it here because the stage and dot theories to be discussed later contain formally similar apparatus.

A commitment of all the theories that follow will be the possibility of analyzing tenses into a tenseless language—tense operators such as 'was' and 'will be' in terms of generalizations over operators such as 'at t'. For example "Was: ϕ" is true relative to $\langle t, w\rangle$ on I iff for some time t' prior to the time of utterance, "At t': ϕ" is true at $\langle t, w\rangle$ on I. "It has always been that ϕ" is true relative to $\langle t, w\rangle$ on I iff for every time t' prior to the time of utterance, "At t': ϕ" is true at $\langle t, w\rangle$ on I.[20] Assuming this has been accomplished, the challenge for our theories is to provide a semantics for the 'at t' operator. Analogous remarks apply to modal operators such as 'necessarily' and 'possibly', which are to be defined in terms of an 'at w' operator; and (in the case of dot theory), to location operators.

In what follows, we write $|e|^I$ for the semantic value of the expression e. A semantic theory that allowed for indexical phenomena would give a story about how $|e|^I$ is determined by the context in which e is uttered—we should therefore regard I as the assignment of content to expressions *in a given context*. The modal operators are to be defined using a primitive symbol of the framework C^m, which may itself be indexical.[21]

- "$\phi \wedge \psi$" is true at $\langle w,t\rangle$ on I iff
  "ϕ" is true at $\langle w,t\rangle$ on I and "ψ" is true at $\langle w,t\rangle$ on I.

- "$\neg\phi$" is true at $\langle w,t\rangle$ on I iff
  "ϕ" is not true at $\langle w,t\rangle$ on I.

- "$\exists x \phi x$" is true at $\langle w,t\rangle$ iff
  "$\phi(c)$" is true at $\langle w,t\rangle$ on some extension I^* of I to the language $L \cup \{c\}$ where $|c|^{I^*}$ exists at $\langle w,t\rangle$.[22]

[19] The terminology here is intended to recall the distinction between 'substance sortals' and 'characterizing predicates' of Wiggins (1980): but neither Parsons nor I wish to be committed to the Wigginsean understanding of the distinction. Some features carry over: for example, an s-predicate F, will be something such that in order to cease to be F, one must cease to be—a Wigginsean result. But, for example, no kind-essentialism need follow from this. See Parsons (op cit) for further discussion.

[20] See Parsons (op cit, appendix) for a sample translation of some temporal vocabulary into a tenseless language. However see Dowty (1979, ch.3) for an argument that, to handle tense and aspect in English, we need not only relativization to instants, but relativization to intervals of time. This is briefly discussed below.

[21] Notice also that I use the quasi-substitutional way of handling quantifiers found in Benson Mates. In an already complex presentation, this allows us to drop at least one of the respects in which truth is relativized.

[22] The quantifier is 'ontologically loaded' in the sense that it only ranges over objects *at one time*. I see no reason why there should not be tenseless and 'possibilist' quantification where this restriction is lifted (it is easy to see how the clauses would be altered. Likewise, one can imagine non-*de re* intensional operators which do not invoke relations. We may be using such devices when we say "there is some German composer who is now famous". (For an alternative strategy, see (Cresswell, 2004)).

- "At w', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t\rangle$ on I iff "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w',t\rangle$ on some extension I^* of I to the language $L^* = L\cup\{c_1^*,\ldots,c_n^*\}$ where $C^m(|c_i|^I,|c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at $\langle w',t\rangle$.

- "At t', ϕ" is true at $\langle w,t\rangle$ on I iff "ϕ" is true relative to $\langle w,t'\rangle$.

- "n is F" is true relative to $\langle w,t\rangle$ on I iff either:
  (1) F is a s-predicate and $|$"n"$|^I$ is a member of $|$"is F"$|^I$; or
  (2) F is a c-predicate and some t-stage of $|$"n"$|^I$ is a member of $|$"is F"$|^I$.[23]

Worm theory reflects the perdurance view. Perdurance says that ordinary objects—things like rabbits—are space time worms. Correspondingly, it is this kind of object that 'rabbit' or 'gavagai' divides its reference over.[24] Perdurance theories say that a thing alters by having successive temporal parts that instantiate different properties—and this is reflected in worm theory by the account of c-predication, whereby a thing satisfies 'is running' at t, and fails to satisfy it at t', by having temporal parts at those times that respectively do and do not fall under the extension of 'is running'.

Stage theory

Stage theoretic semantics reflects the *exdurance* theory of persistence advocated by Sider (1996a, 2001) and Hawley (2001).

> According to the [exdurance] theory, ordinary persisting objects are stages that persist ... by having distinct stage counterparts at other times. [Exdurance] says that in the afternoon when I find my bent candle on the shelf, the candle is the bent-stage coexisting with me then, but that stage persisted from before (in the relevant sense) by virtue of having a (straight) counterpart stage on the shelf in the morning. ... Although on this view ordinary objects are stages and so (strictly speaking) only exist momentarily, they can nonetheless persist by virtue of having counterpart antecedent and/or successor stages.
>
> (Haslanger, 2003, p.318)[25]

The key features of the exdurance theory are that ordinary objects are strictly speaking momentary stages, persisting vicariously through their counterparts located elsewhere in time; and that things change their properties through having counterpart-stages at successive times instantiating different properties.

To set up stage theoretic semantics to reflect the exdurance theory of persistence and alteration, we need an additional primitive: the two-place relation C^t ("is a temporal counterpart of"). This holds between temporal slices. Intuitively, $C^t(x,y)$ will hold if x and y are stages of the same persisting object—in the case at hand, if they are stages of the same rabbit.

- "$\phi\wedge\psi$" is true at $\langle w,t\rangle$ on I iff "ϕ" is true at $\langle w,t\rangle$ on I and "ψ" is true at $\langle w,t\rangle$ on I.

- "$\neg\phi$" is true at $\langle w,t\rangle$ on I iff "ϕ" is not true at $\langle w,t\rangle$ on I.

[23] More generally, if we have a polyadic predicate R, with k s-places and l c-places then "$Rn_1,\ldots n_k,m_1\ldots m_l$" is true relative to $\langle w,t\rangle$ on I iff $\langle n_1,\ldots,n_k,m_1',\ldots,m_l'\rangle$ is a member of $|$"is F"$|^I$, where m_i' is some t-stage of m_i, for each i.

Note that we will now have a *variety* of relations with claim to be the identity relation $a=b$, according to whether a and b are regarded as of sort s or sort c. This leads to real differences in the context of counting.

[24] Or at least, it will do so so long as we regard 'gavagai' as what Wiggins would call a 'pure sortal', and not, for example, a phase of the underlying sortal 'organism'. In the former case, 'gavagai' will be a s-predicate, holding of all rabbit-worms; in the latter case, 'gavagai' will be divide its reference over rabbit-stages. See Parsons (op cit) and the references therein.

[25] Haslanger (2003) uses 'exdurance' and 'stage theory' interchangeably. I find it useful to regiment the terminology to distinguish metaphysical from semantic theories, and have altered the above quote accordingly.

- "$\exists x\phi x$" is true at $\langle w,t\rangle$ iff
  "$\phi(c)$" is true at $\langle w,t\rangle$ on some extension I^* of I to the language $L\cup\{c\}$ where $|c|^{I^*}$ exists at $\langle w,t\rangle$.
- "At w', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t\rangle$ on I iff
  "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w',t\rangle$ on some extension I^* of I to the language $L^*=L\cup\{c_1^*,\ldots,c_n^*\}$ where $C^m(|c_i|^I, |c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at $\langle w',t\rangle$.
- "At t', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t\rangle$ on I iff
  "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w,t'\rangle$ on some extension I^* of I to the language $L^*=L\cup\{c_1^*,\ldots,c_n^*\}$ and $C^t(|c_i|^I, |c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at$\langle w,t'\rangle$.
- "c is F" is true relative to $\langle w,t\rangle$ on I iff
  $|$"c"$|^I \in |$"F"$|^I$.[26]

This stage theory reflects the exdurance view. Exdurance says that ordinary objects—things like rabbits—are momentary stages. Correspondingly, 'rabbit' or 'gavagai' divides its reference over this kind of object. Exdurance theories say that a thing alters in virtue of distinct counterparts having distinct properties —and indeed, we find that "a rabbit was running and is now sitting still" will be true iff some present rabbit-stage that falls under "sitting still" has a past temporal counterpart that falls under "is running".

Dot theory

When it comes to undetached rabbit parts, there is no corresponding theory in the literature to draw on. One can see what such a theory would look like, however: it would be a theory of extension through space that paralleled the exdurance theory of persistence through time. On such a view, a rabbit that is partially at a position p would extend through space, not by having parts located at other places, but through having *counterparts* at those positions.

We can modify the stage theoretic ideas to get a corresponding semantic theory, which I shall call 'dot theory'.[27] As before, we have primitive counterpart relations C^m, C^t, and in addition a relation C^p relating one object to another at a single time. In the case at hand, the intended relation will hold of the pair x,y if they are parts of the same rabbit (stage).

- "$\phi\wedge\psi$" is true at $\langle w,t,p\rangle$ on I iff "ϕ" is true at $\langle w,t,p\rangle$ on I and "ψ" is true at $\langle w,t,p\rangle$ on I.
- "$\neg\phi$" is true at $\langle w,t,p\rangle$ on I iff "ϕ" is not true at $\langle w,t,p\rangle$ on I.

[26] The generalization to polyadic predicates is not immediate. In particular, we would like to be able to handle *crosstemporal relations*—for example "William I was the ancestor of Elizabeth II". The problem is that there may be no single instant where both relata have counterparts.

Nothing in the above requires that the objects in the extension of a predicate at t need themselves to exist at t. Indeed, there are good reasons for resisting this principle. Consider, for example, the predicate "is famous". It is natural to think that Beethoven satisfies "is famous" relative to the present time (and fails to satisfy it, for example, relative to a post-nuclear world where classical music has been forgotten). But Beethoven does not exist at the present time. (This example is from Cresswell (2004) who takes it to motivate the construction of ersatz counterparts for individuals in worlds and times where they do not exist. The above seems to me a more attractive handling of the case.)

If we adopt this setup, then there is no immediate objection to having $\langle William, Elizabeth\rangle$ within the extension of 'is an ancestor of' relative to the present moment.

An alternative treatment would appeal to relativization to intervals, and say that pair is in the extension of 'is an ancestor' relative to the 2nd millennium. See p. 4.5.1 below.

[27] Wolfgang Schwarz (2005a), discusses an 'atomistic' theory he calls 'general counterpart theory'. The version presented derives from an early version of Schwarz's paper though with some variations to parallel more exactly the counterpart theoretic paraphrases of Lewis (1968). Schwarz has subsequently reformulated his theory in ways that are less congenial to my present purpose.

- "$\exists x \phi x$" is true at $\langle w,t,p\rangle$ iff
  "$\phi(c)$" is true on some extension I^* of I to the language $L \cup \{c\}$ on which $|c|^{I^*}$ exists at $\langle w,t,p\rangle$.

- "At w', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t,p\rangle$ on I iff
  "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w',t,p\rangle$ on some extension I^* of I to the language $L^* = L \cup \{c_1^*,\ldots,c_n^*\}$ such that $C^m(|c_i|^I, |c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at $\langle w',t\rangle$.

- "At t', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t,p\rangle$ on I iff
  "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w,t',p\rangle$ on some extension I^* of I to the language $L^* = L \cup \{c_1^*,\ldots,c_n^*\}$ where $C^t(|c_i|^I, |c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at $\langle w,t'\rangle$.

- "At p', $\phi(c_1,\ldots,c_n)$" is true at $\langle w,t,p\rangle$ on I iff
  "$\phi(c_1^*,\ldots,c_n^*)$" is true relative to $\langle w,t,p'\rangle$ on some extension I^* of I to the language $L^* = L \cup \{c_1^*,\ldots,c_n^*\}$ where for each i, $|c_i^*|^{I^*}$ where $C^p(|c_i|^I, |c_i^*|^{I^*})$ and $|c_i^*|^{I^*}$ exists at p'.

- " c is F" is true relative to $\langle w,t,p\rangle$ iff
  $|\text{“c”}|^I \in |\text{“}F\text{”}|^I$[28]

We can now let the extension of "Rabbit" or "Gavagai" be all and only *mereologically atomic* undetached parts of a rabbit.[29] An atomic part falling under 'gavagai' will extend in space vicariously through having p-counterparts in other places. It will change its properties through space through (for example) having white counterparts in one place; and counterparts in another place that are black.[30]

The primitives

One of the most striking things about the move from worm to stage to dot semantics is the need to appeal to additional primitives in each case. For worm theory, we appealed to the counterpart relation C^m (and recall, this was inessential—we could have used some other treatment of modality if preferred.). For stage theory we have in addition the temporal counterpart relation C^t, and for dot theory, we also have the positional-counterpart relation C^p.

The modal counterpart relation as *Lewis* understands it, is a contextually inconstant relation, so that statements about necessary properties (and *de re* modal predication in general) become highly sensitive to peculiarities of conversational context.[31] Lewis thinks of the counterpart relation as being fixed by facts about the *similarity* of objects to others. Although facts about similarity are not treated as in any way subjective by Lewis, *which respects* of similarity are called upon in a given case will depend on the demands of the conversation.[32]

The *formal* appeal to 'counterpart' relations does not commit us to anything substantive about the *nature* of that relation. For example, it as yet says nothing about whether we can further analyze it (as Lewis analyzes modal counterparthood in terms of similarity); or whether we should treat it as primitive. Even in the modal case, compatibly with all that has been said, which objects are the modal counterpart of one another might depend on their instantiating special 'haecceities' (individual essences). This could lead to a treatment of modal predication very different from Lewis', though sharing the same formal setting.

[28] To deal with cross-temporal and spatial relations, similar moves to that used in stage theoretic setup will be needed.

[29] To get arbitrary undetached parts within the extension of 'gavagai', as on Quine's original proposal, we would need to combine elements of the dot theory with elements of the worm theory. I will not give details here.

[30] For purposes of exposition, we indulge in the fiction that mereological atoms are intrinsically coloured.

[31] Lewis (1968), Lewis (1986c, §4.5)

[32] Lewis (1983a). Perhaps the most extreme example of Lewis endorsing context-sensitivity is to be found in "Things *qua* truthmakers" (2003).

Appreciating the distinction between the flexibility of the *formal* apparatus of counterpart theory and the particular doctrines endorsed by its inventor is important when thinking of the stage and dot theoretic versions. It may help to think of the stage-theorist's "temporal counterpart relation" as a relation of temporal *unity*. Every participant in the current debate owes a story about how stages hang together to form continuing objects, even if this only amounts to taking the notion of a 'natural united' object as primitive. This 'hanging together' is exactly what C^t expresses (cf Hawley, 2001, ch.3).[33]

We can categorize accounts of the counterpart relations in numerous cross-cutting ways: (1) It could be taken as primitive, or an analysis could be offered. (2) It could be *constant*—a single relation-in-extension no matter what the context is; or *inconstant*—which relation-in-extension is picked out by C^t might vary.[34] (3) Further, *grades* of inconstancy might arise. One might hold that there are only two temporal counterpart relations; say counterpart-*qua*-body and counterpart-*qua*-person.[35] Or, like Lewis, one might think of endless counterpart relations, whose selection is highly dependent on context. (4) On an inconstant view, one might hold that distinct counterpart relations are associated with each sortal predicate (so that sortal predicates 'carry with them' criteria of diachronic persistence)[36]; or one might not endorse such a connection. (5) One might think of the counterpart relations as subjectively constituted—for example, in terms of the classificatory dispositions of agents; or objectively constituted—for example, in terms of objective similarity (in the modal case)[37] or constitutive causal relations (in the temporal case)[38]. Of course, the modal, temporal and part counterpart relations may require analysis in different ways.

My concern here has been solely to set up the formal framework, and I intend to have presented it in a way that is compatible with any of these positions. In assessing whether the various proposals can be cashed out in a way that meets the minimal constraint of fit with the patterns of assent and dissent that form the interpretationist's data, we shall occasionally fill out the proposal in one direction or another, but in the present context, neutrality is a virtue, and one that I shall seek to preserve as far as possible.

[33] On a role for similarity in uniting successive stages, see Lewis (1976). It is clear that *similarity* will not find much use in an account of the p-counterpart relation involved in dot theories.

[34] As Parsons (2005a) notes (in a slightly different context), the kind of views on diachronic identity espoused by Ayers (1974) and Armstrong (1980) may lead to a constant temporal unity relation.

[35] A picture that suits this case is one where we appeal to a small number of temporal unity relations, each paired with a specific natural kind. Hawley (2001, ch.3, esp. p.70).

[36] See Hawley (2001, §5.5)

[37] See Lewis (1986c, p.254)

[38] See Armstrong (1980)

4.5 Objections

In the next few sections, we consider objections to the argument for division inscrutability based on the three semantic frameworks just sketched. These are based on the concerns of Evans and Fodor, reviewed earlier (we also discuss briefly a common objection to stage theory—that it has problems with counting statements). We first cover Evans' concerns about how stage-division handles tense ((C) on 72, above). Our formulation of the Quinean proposal in terms of stage theory now comes into its own: stage theory includes (or can be straightforwardly extended to include) the temporal operators needed to make sense of the semantics of tense and aspect. We then move to investigate the predication-based objections to URP-division ((A) and (B) on 71, above). I shall present a general feature of the use of counterpart theory—the phenomenon of 'inconstant' predication, and point to potential instances in both worm theory (potential modally inconstant predication) and stage theory (potentially temporally inconstant predication). I then show how the Evans and Fodor objections to URP-division surface, within the dot theory, as instances of this general phenomenon.

4.5.1 Tense and aspect within stage theory

Worm, stage and dot theory each include temporal operators in terms of which tenses can be defined. Evans' concerns about the ability of the stage-view to deal with tensed attributions can thus be answered in the most satisfying way: by appealing to a general treatment of tense.[39]

The more general concern, however, may still be a good one. Evans highlights the need for a systematic account of tenses in English. Though *tenses* (strictly construed) can be handled, it is not so clear that other forms of temporal relativization will be. It is not clear at first glance whether the stage view can handle temporal *aspect*.[40] More particularly, as currently formulated, stage theory allows relativization to *instants* in time only. However, this may be inadequate to provide semantics for natural language. Dowty (1979, ch.4.) claims that to handle, for example, the English progressive ('John was crossing the road') we need to appeal to relativization to *intervals*.[41]

In fact, we can make a case for relativization to intervals more directly. Consider 'John read *War and Peace* yesterday' (McCawley, 1980, p.345). For this to be true, it is not enough that there be a time instant yesterday at which John was reading the book (for that is compatible with him not finishing it), nor that be a time instant yesterday at which John finished reading the book (since that would be consistent with his having only read the final page yesterday). Various more complex analyses might be tried, but, having considered several such analyses, McCawley writes:

> I would like to propose that not only points in time but also intervals figure in the logical structures of sentences and that examples such as [those above] all involve time intervals. ['John read *War and Peace* yesterday'] will then not say that there is a past time at which John read *War and Peace* but that there is a past time interval such that he read it on that interval.

[39] Essentially, the flaw in Evans' original argument lies in failing to distinguish what Sider (1996a) calls *de re* and *de dicto* temporal predications: $\exists x\mathrm{WAS}Fx$ vs. $\mathrm{WAS}\exists xFx$. This is effectively the objection to Evans offered in Wright (1997) and Richard (1997).

[40] Josh Parsons mentioned this puzzle to me as a known problem for the Sider/Hawley stage theory of persistence. However, I have not been able to locate a reference within the literature on stage theory.

[41] . The basic idea is that 'John is crossing the road' is true relative to a time T if there is a interval T' including T relative to which 'John crossed the road' holds. This needs modification, though, to deal with the so-called 'imperfective paradox': 'John was crossing the road' can be true in situations where he never finished crossing (e.g. because he was run over). See Dowty (op cit) for discussion.

(McCawley, 1980, p.345)[42]

Let us suppose that the analysis of tense and aspect involves irreducible interval-relativization, as McCawley and Dowty urge. Still, division over rabbit stages seems in good order, since we can simply re-interpret the stage semantics of p.75 above, so that the temporal indices range over intervals rather than instants. Formally, everything else remains the same—for example, the extension assigned to a predicate at an interval will be a set of instantaneous temporal parts. Relative to the interval during which John read *War and Peace* (say, 1am to 11pm) each John-stage within the interval will fall under "reads *War and Peace*", whereas relative to a time when he only got through half of it, no John-stage will fall under that predicate.[43]

Digression: counting rabbits

One common objection to stage theory is that it delivers the wrong results for counting statements.[44] The intuitive objection is that, since the reference of 'rabbit' is divided over infinitely many instantaneous rabbit stages, it will say that there are infinitely many rabbits in the hutch during a period in which common sense tells us that there is only one.

A standard Quinean response to such moves would be to re-interpret the 'apparatus of individuation': to declare that under the stage-hypothesis, we do not count by identity but by some *ersatz* relation.[45] That is, instead of characterizing the extension of 'is identical to' through numerical identity, we characterize it using the counterpart-hood relation. Writing 'I' for short, we let 'xIy' be true relative $\langle w,T\rangle$ iff $C^t(x,y)$ holds. Under this interpretation, we will not get errors of counting, since if (intuitively) Mopsy alone is in the hutch during Tuesday, all the rabbits in the hutch during that period are I-related.

Sider (1996a) declares it a virtue of stage theory that one does *not* have to do this, but can stick with the interpretation of identity as $=$. However, he does think that re-interpretation is needed to get sensible results when, for example, counting the number of rabbits in a hutch over an extended period of time.[46]

For the Quinean, however, counting by I seems entirely unproblematic. Moreover, it is not obvious that we are *re-interpreting* identity using I. After all, I only relates things that are, will be or were numerically identical.[47] I conclude that the Quinean, at least, has no worries about counting.

(This analysis also deliver interesting results in so-called 'fission' cases (Sider, 1996b; Lewis, 1976). If Mopsy undergoes fission, splitting amoeba-like to become Mopsy1 and Mopsy2, then Mopsy1 and Mopsy2 will witness the truth of 'there are at least two rabbits in the hutch during Tuesday' (there are two rabbits—a Mopsy-1-stage and a Mopsy-2-stage who are not I-related). The interesting result, however, is that we may get 'there is exactly one rabbit in the hutch during Tuesday' coming out true as

[42] See also Dowty (1979, ch.3.).

[43] Related predicates, such as the progressive 'is reading *War and Peace*' will hold of John-stages at instants during that day, given an appropriate treatment of the progressive (Dowty, 1979, ch.4.).

[44] See in particular Sider (1996a), who despite defending stage theory is particularly concerned about this point.

[45] See Field (1974) for an example of this Quinean approach. Sider (1996a) endorses re-interpretation of identity for particular 'counting' contexts.

[46] It is not obvious that counting by $=$ itself will straightforwardly deliver bad results, once we factor in the tense and aspect within counting statements. However, I do think that, particularly taking into account the interval quantification just mentioned, it faces severe problems. In particular, we need to check that a proposal handles the following case correctly: Flopsy is in a hutch during the early part of Tuesday; then taken out and destroyed. Mopsy is created, occupies the hutch during the late part of Tuesday, and then is destroyed. On Wednesday, there should be a reading of 'there were two rabbits in the hutch during Tuesday' which is true. The challenge for one who wishes to 'count by $=$' is to show how the formulation of counting that handles this case correctly can also deliver the intuitively correct results when the reference of 'rabbit' is divided over stages.

[47] In particular, this is not the 'counting by rabbit-worms' that Sider (1996a) appeals to. In the famous case, 'the Statue=the Clay' will be true, for example, in virtue of their present stages being identical (and so counterparts), whereas they may well be distinct worms. A better description would be 'counting by ersatz identity', since we can regard counterpart-hood as the stage theorist's substitute for diachronic identity.

well. The point is that there is a rabbit within that hour (a pre-fission stage of Mopsy) which is I-related to every rabbit in the hutch within the day. There are two familiar ways of handling 'there is exactly one F' within first-order logic:

> there is at least one F and it is not the case that there are at least two Fs
>
> there is at least one F and every other F is identical to it

On the former reading the Mopsy-fission scenario will not make-true 'there is exactly one rabbit in the hutch'. But on the latter reading, the Mopsy-fission will make-true *both* 'there is exactly one rabbit in the hutch' and 'there are exactly two rabbits in the hutch'. I find this result quite appealing—an apt reflection of the confusing nature of counting in fission cases.[48])

4.5.2 Predication and compounding

We now turn to the Evans and Fodor objections (A) and (B) (given on page 71ff above.). The natural proposal is to let anything which is (intuitively) an atomic part of an F fall under 'F' on the dot-semantics. As anticipated, *prima facie* problems with compounding arise: we will indeed have objects that fall under 'Rabbit' and 'White' and 'Ear' in the presence of a rabbit with a white ear—just take any simple part of the ear.

Rather than tackle the problem directly, I want to outline a range of parallel cases within the worm and stage frameworks, and describe how advocates of those positions are likely to react to the challenges. I will then look at how the analogous responses would work within the dot semantics.

Inconstancy within stage and worm theory

Consider stage theory, as developed by Sider (1996a). The proposal is to let any stage of something that is intuitively an F fall under the extension of 'F'. The temporal counterpart relation is something that unites person-stages of a single person: Sider takes it to be a matter of psychological connectedness. In the case of other kinds of objects, different kinds of temporal 'unity' relations would be needed. Our worries will arise when two objects which *share a stage* nevertheless call for extensionally distinct counterpart relations.

Let us illustrate this with a famous case. Consider a lump of clay that has existed for millions of years, and which was formed into a statue 35 years ago. Many wish to maintain that the statue *came into existence* when the clay was formed into the shape of a statue—it is an object in its own right, not merely a temporary property of the lump of clay.

From the stage theorist's perspective, this presents a dilemma. Does the temporal counterpart relation relate the present statue/clay stage to a stage of the piece of clay before it was formed into a statue? Suppose it does: then the statue pre-existed its sculpting, in virtue of having a temporal counterpart before that event. Suppose it does not: then the piece of clay has no counterparts before the sculpting, so came into existence at that point also.

Sider's analysis of the statue/clay case is to admit *two* temporal counterpart relations: statue-counterparthood and lump-of-clay-counterparthood. The latter relates the statue/clay stage to entities pre-existing the sculpting event; the former does not. Which relation is designated by our relation C^t is a matter for

[48] The proof that the two readings of 'there is exactly one F' are equivalent relies on the Euclidean property of the identity relation: and it is exactly this that fails when numerical identity is replaced by the *ersatz* identity I. Thanks to Andy McGonigal for this point.

A similar treatment of fission cases will arise within worm theory, if the name-positions flanking the identity sign are treated as c-predicates (cf. p.74, above.). Thanks to Wolfgang Schwarz for suggesting this.

context to decide. When we ask about the creation of the statue, the former is invoked; when we ask about the pre-existence of the clay, the latter is invoked.

What we have described thus far is sufficient to deal with questions phrased in terms of quantification (existence) and temporal operators. Using such devices we can ask about whether the statue existed 36 years ago, and get one answer; and we can ask about whether the clay existed 36 years ago and get a different answer. This is compatible with maintaining that the statue and the clay are identical.

What I now want to highlight is that this does *not* solve all the problems that we need to ask. For there are predicates that depend on the distribution of properties over the course of an object's history. Such 'historical' predicates include, paradigmatically 'is exactly 35 years old' and 'is millions of years old'. Again, a dilemma emerges: is it true to say "there is a million year old statue present"? Presumably not, given that the lump of clay was formed into a statue only 35 years ago. Nevertheless, since the lump of clay stage is identical to the statue stage, if one is within the extension of "millions of years old" the other must be too. We *do* want "there is a lump of clay present that is millions of years old" to come out as true: so we are pressured towards admitting the former sentence as true. Given that exactly similar remarks could be made in favour of placing the stage inside the extension of "is exactly 35 years old" we are in danger of declaring true "there is something that is both exactly 35 years old and is millions of years old".

One option here is to appeal to *paraphrase* so as to reduce the problem to one already solved. The basic idea is to map

> n is (at least) 35 years old

to

> Throughout the past 35 years, n has existed.

As before, inconstancy of the counterpart relation directly impacts here, given the way that temporal operators are defined.

The fundamental objection in the context of finding a possible *semantic analysis* of our language, is that the free use of paraphrase to turn predicates into operators looks illegitimate. If we are asking for a semantics for a language with a fixed syntax, we need some other device: for we want an interpretation of the *predicate* 'is 35 years old'. Pointing to an operator that systematically corresponds to it is not to deliver this.[49] The resolution is close to hand, however. What we must maintain is that not only the counterpart relation, but a range of related predicates are inconstant or indexical. 'Historical' predicates such as 'is exactly 35 years old' are paradigmatic examples of this class. Indeed, the two indexicals are related in a natural way: 'is exactly 35 years old' will hold of a stage at *t* iff the sum of the temporal counterparts of that stage existing earlier than *t* measures 35 years. Notice that we *use* the counterpart relation in specifying the property. Hence any indexicality characteristic of the counterpart relation will infect the extension of the predicate.[50] We get:

> The statue is the clay
>
> The statue is exactly 35 years old
>
> The clay is millions of years old.

To explain the difference, we point to changes in context: the latter two invoke different counterpart relations (statue-counterparthood and clay counterparthood respectively) which in turn changes the extension of the relevant predicates.

[49] What we would need to make a principled case is some independently motivated transformation or generative component in the semantics that would derive the surface predicate from underlying operators. Cf. Lewis (1970a) and Dowty (1979, ch.1.).

[50] Compare the 'indexical' response to Evans suggested by Wright (1997, p.410).

The residual question

Our resolution of the puzzles over inconstant predication make heavy use of contextually varying counterpart relations. This leaves a residual worry. For simplicity, I set up the semantics within a single context, and assumed that this context would determine a unique counterpart relation. It is natural to think, however, that different counterpart relations *can* be invoked by different parts of the same sentence. Thus (naming the statue 'Goliath' and the clay 'Lump') it is natural to think that

Goliath is exactly 35-years old and Lump is millions of years old (*)

should come out true. For this to be the case, we need to allow *different* counterpart relations to operate in the two conjuncts. Two worries then emerge: How are we to think of this case? What prevents us from existentially generalizing to get the problems back once again?

On the first point, we should first note that we have independent reason to think that context-change within a sentence can take place. Consider the utterance "Now is not... now" Since the second utterance of 'now' takes place at a later time than the first, there is a natural reading of the sentence on which it expresses a truth. On the other hand, relative to any single context, it expresses a contradiction. I suggest the following view of how token utterances get assigned truth-conditions by a semantics. First, each expression in the utterance has its own context. The first component to semantic theory (what Kaplan (1989b) calls 'character') will assign to each component a content, depending on its own unique context. Indexical terms such as 'I' and 'now' are assigned a referent, indexical predicates are assigned an extension, and so on. Once this is fixed, the second component of semantic theory kicks in. This tells us how the referents and extensions assigned to the various expressions combine to determine the overall truth-conditions for the sentence. The various semantic theories that we give above are each candidates for this second component. Which extension the predicates have, or what precise relation the primitive terms of the theory express, may indeed vary depending on the context at hand, even within the same sentence.[51]

The second worry seems to me the most serious challenge. What prevents the worrying existential generalization:

There is something that is both exactly 35 years old and is millions of years old?

We have been given no reason to think that context cannot change in ways that render true an utterance of such a sentence. The case is particularly pressing given our verdict that the witnessing statement ($*$), above, is unproblematic.

There are two responses. The *way of resistance* is to find some principled constraint on change of context which prevents the generalization from coming out true. It is hard to see a non *ad hoc* route for this. Better, then, is the *way of concession*. This is to accept that there is no principled reason that a context cannot be found with respect to which the existential generalization is true; but to insist that such contexts do not normally arise, so that *standardly* the existentially generalized sentence expresses something false. One would then hope to *explain away* intuitions that the statement is false, on the grounds of its typical (though not inevitable) falsity.

I advise the stage theorist to take the way of concession. One should try to explain intuitive resistance to the bare existential by noting that for it to be true, context would need to invoke statue-counterparthood

[51] For more examples of how context change within a sentence can be significant, see Lewis (1979c). The account of belief reports in Stalnaker (1999a) also requires context change. It seems to me that the need to distinguish *contextual determination* of content, involving a variety of contexts, from the calculation of truth-conditions of a given utterance, once content of its parts have been fixed, undermines Lewis (1980)'s suggestion that we could do semantics entirely in terms of a single binary functions from index and context to truth-values. It gives a principled reason for discerning a significant level of 'content' within the overall 'semantic value'.

for the first part of the sentence, and clay-counterparthood in the second. However, there are no prompts for such change in the sentence. If we do add such prompts, we get something that sounds (to my ears at least) acceptable:

> there is something that is exactly 35 years old (*qua* statue) and millions of years old (*qua* lump of clay).

Analogues in the worm case

I have just explored the way that Sider handles statue/clay cases within stage theory. The key was to diagnose inconstancy in the counterpart relation, and to extend this to relevant 'inconstant' predication. I now want to briefly sketch the analogous case within worm theory, before turning to the Evans/Fodor objections to Quinean inscrutability.

Worm theory, as we have set it up, puts fusions of temporal stages of Fs into the extension of 'F', when F is a s-predicate, and puts G-ing stages into the extension of 'G', when G is a c-predicate. It straightforwardly handles the whole variety of cases we have hitherto considered.[52] Nevertheless, analogous phenomena *do* arise, given modal counterpart theory. It is not unreasonable to hold that the statue is *essentially* a statue.[53] Equally, it is not unreasonable to hold that the lump of clay is *essentially* made of clay, but might never have been made into a statue.[54] On the other hand, if the statue and the clay were created and destroyed at the same time (unlike Goliath and Lump) the statue and the clay will be the same space-time worm.

Lewis (1986c, §4.5) explicitly admits this kind of case, by making allowance for inconstant modal counterpart relations (indeed, his is the model that Sider follows in developing the stage view). Lewis' thought is that on some standards of similarity, the statue/clay worm can be similar enough to the statue to count as its counterpart in another possible situation; but on other standards of similarity no non-clay can be a case in point, but statue-hood is irrelevant. We can retrace the steps described earlier for the clay case: pointing to possible paraphrase in terms of modal operators to which the modal counterpart relations are directly relevant (e.g. 'Necessarily, it is a statue'); and then looking at an interpretation of the predicate 'is essentially a statue' specified in terms of counterpart relations, so that it would inherit the indexicality of the latter.

Again, there would be problematic existential generalizations:

> there is something that is both essentially a statue; and is essentially made of clay, but might never have been a statue

Again, we try to allay worries by noting how odd the context-change involved would have to be to render this true; and also, perhaps, the acceptability of versions where contextual prompts such as *qua statue* are introduced.

The Evans and Fodor cases revisited

What now of the dot-semantics, and the Evans and Fodor objections? I want to urge that we see the Evans/Fodor objections as the surfacing of the phenomenon of inconstancy found above. Consider first Evans' challenge: of what objects does 'white' hold? The underlying problem here is that whether or not a rabbit is white depends on the overall distribution of whiteness in its fur. Like the historical predicates that were problematic for stage theory, 'spatialized' predicates will pose challenges for dot theory.

[52] This has the rather disarming implication that, strictly speaking, nothing is both white and a rabbit. Of course, the semantics of predicates are set up in such a way that 'everything that is a white rabbit is a rabbit' will come out true.

[53] Gibbard (1975)

[54] Kripke (1980, ch.2)

The solution is to characterize the extension of predicates informally via the contextually salient counterpart relation:

> 'is white' applies to *a* iff the fusion of *a* and its p-counterparts have a white outer surface.

When 'rabbit-counterparthood' is salient, all atomic parts of a single rabbit are counterparts of each other. In the presence of a white rabbit, the condition will be met by any part of a white rabbit. It is not met by any part of a black rabbit with a white ear. It is not met even if there is an atomic part *A* (part of the ear of the rabbit) which falls under 'rabbit' and which is itself intrinsically white in colour.

Each such sortal will have to deliver its own counterpart relation. For example, there will be ear-counterparthood, under which *x* and *y* will be counterparts iff they are both parts of the same rabbit ear. In the scenario sketched above, the same object *A* which did not fall under 'white' under the rabbit-counterparthood relation, will fall under 'white' under ear-counterparthood.

I hope it is clear that the above is just the analogue of the treatment of inconstant predicates in the modal and stage settings: though now almost every predicate is inconstant. We can expect an analogue of the odd existential generalizations found earlier. In the current setting these generalizations are something like:

> There is something that is both white all over and mostly black

For *A*, above, is white all over *qua* part of a white-all-over ear; and mostly black, *qua* part of a mostly black rabbit. Again, our initial discomfort might be explicable given the changes of context that must occur to render the odd-sounding sentence true; and might be disarmed if we find the '*qua*' glosses moderately acceptable.

More disturbing is the Fodorian challenge:

> There is something that is both a rabbit and an ear

Prima facie, we have here a choice between two uncomfortable options. The first is to regard 'rabbit' and 'ear' as constant predicates, applying invariantly to atomic undetached rabbit parts and atomic undetached ear parts. That involves regarding the above as on all fours with statements in the stage/modal cases such as 'there is something that is both a statue and a lump of clay'. Such statements were unproblematic in those contexts: but the Fodorian analogue is far less comfortable.

The other approach is to say that sortal predicates, as well as adjectives such as 'white', are inconstant: we would then give characterizations of the extension of sortals such as:

> '*x* is a rabbit' is satisfied by *A* iff the fusion of *A* and its counterparts makes up a rabbit.

Under the ear-counterpart relation, the extension of this predicate will be empty; under the rabbit-counterpart relation, all the atomic undetached rabbit parts will fall within it. Of course, this doesn't mean that there are any ordinary contexts in which 'there are no rabbits' would be true: for such sentences *ipso facto* make salient rabbit-counterparthood. The view does allow us to regard the Fodorian sentence above as the direct analogue of the cases familiar from stage and worm views: just as in those cases, the diagnosis will be that the sentence can only express a truth if there is a context-change occurring in the middle of it. At first glance, this hardly looks to be an improvement on the previous option, since the predicates involved are exactly the ones that generate the counterpart relations, it is hard to see why the sentence would be unacceptable.

What one would need to resist Fodor's objection, then, is either a complex story about how the context-change needed to render true the relevant generalization is for some reason not available; or else a healthy propensity to bite bullets. Suppose we tug on a rabbit's ear, and say: "this is a white ear; but it is also a mostly black rabbit". In context, the statement seems fine. We would probably regard it as a pun[55]—but for the dot theorist, it would express the sober truth.

Some philosophers describe the part-whole relation as partial identity: the part being partially identical to the whole. The dot theorist takes this in the most literal way: the part (the ear) is identical with the whole (the rabbit). We can of course introduce non-standard interpretation of 'identity' (e.g. 'whole identity'[56]) such that the ear and the rabbit would not be *wholly identical*. However, this will not finesse the Fodorian problem we have just seen, for the problematic sentences make no use of identity, but only quantification and predication. However, it will give us the terminology with which to make some sort of sense of the result.

The Fodorian existential generalization seems a serious problem for the dot-inscrutabilist—it certainly seems unfaithful to ordinary patterns of assent and dissent. Progress has nevertheless been made: we have seen how Evans' worries can be dealt with in ways exactly analogous to corresponding problems for stage and worm theories. We have also seen exactly why the Fodorian generalizations arise: cases where one sortal applies to parts of things falling under another will generate parallel existential generalizations and identities on all three accounts. Only in the dot case do they seem intuitively repugnant. What I take this to show is that the Fodorian phenomena are not symptoms of a wider malaise for dot-theory—they are the most problematic features of a generally successful account. Perhaps more importantly, we have seen that we will not be able to find corresponding problems with the stage view: the analogous sentences (e.g. 'there is something that is both a lump of clay and a statue') are unproblematic. Even if the Fodor objections disrupts the 'undetached rabbit part' version of putative division inscrutability, stage/worm inscrutability is still live.

[55] That is, one would assume that the anaphoric reference cannot be taken seriously—we would ordinarily assume that we need 'this' and 'it' referring to different entities for the sentence to come out true.

[56] E.g. we say that a is wholly identical to b iff every counterpart of a is identical to some counterpart of b, and vice versa.

4.6 Conclusion

Quine's "argument from below" falls into two halves: inscrutability of logical form; and division inscrutability. Quine proposes three alternatives in the latter case, which we have cashed out systematically in worm theory, stage theory and dot theory. Dot theory is clearly of a more dubious standing than the other two: Fodor's examples, at least, threaten to illustrate seriously counterintuitive results. Even this is not decisive, especially once we realize that Fodor-style examples are localized phenomenon.

Let us focus, however, on the most plausible case for division inscrutability: stage and worm semantics—regarding 'rabbit' as dividing its reference over perduring objects, or alternatively over instantaneous stages that persist vicariously through temporal counterparts at other times. I have argued that, for the fragment of language we have been considering, each does as well as the other in matching patterns of assent and dissent—the kind of data that interpretationism takes to settle semantic facts.

In later chapters, we will be looking at additional constraints that the interpretationist may make on the evaluation of semantic theories. The most important such constraint is simplicity. It is worth considering, therefore, whether considerations of simplicity could resolve putative division inscrutability.[57]

It is notable that stage theory features an extra primitive, by comparison to worm theory—the temporal counterpart relation. One might think, on those grounds, that it should be counted a simpler semantic theory. However, this would be to ignore the corresponding complexity of the objects over which worm theory divides reference, by comparison to stage theory. It is a familiar theme that one can give the appearance of simplicity to a semantic theory simply by building relevant information into the domain over which one quantifies: one can, for example, get the power of an intensional setting in a purely extensional way, roughly by quantifying over 'intensional objects'.[58] Everyone owes an account of what makes an ordinary object 'hang together' over time—even if it just taking the notion of a 'naturally unified' object as primitive. This information is encoded within the *objects* that worm theory quantifies over; it is explicit in the temporal unity relation of stage theory. On grounds of simplicity, there is little to choose.

Nor, I think, is there any sense in which one choice is 'the natural' one. This is obscured when we described the options as dividing over rabbits; or over rabbit-stages. This makes it sound like the former is the natural choice. However, when we give a metaphysical explanation of what we take these 'rabbits' to be (certain four dimensional solids, only partially present at a time) it becomes a real issue which, if any, of these candidates are 'truly' rabbits. If what we are interested in is *interpretation* rather than translation—a pairing of words with parts of the world—then there will be a point at which ordinary intuitions no longer have authority. We seem to know a lot, introspectively, about the subject-matter of our thoughts and talk. However, we do not *introspectively* access the physical nature of the things we talk about; and neither should we be thought to have introspective access to the *metaphysical* nature of such things. The point of view I am urging sees our intuitions as silent—or equally bemused—by all the various options we have considered.[59]

Perhaps, once we see the options, there are advantages in choosing one rather than another as a way of *regimenting* our talk of ordinary objects such as rabbits. (The considerations adduced in Sider (1996a) seem to me to be mostly of this nature: Sider even calls them 'philosophers reasons' for preferring a stage theory over a worm theory.) While there may be *normative* reasons for choosing one or the other

[57] In other settings, causal considerations are centre-stage. As noted by Field (1974); McGee (2005a) and others, finding causal relations that would discriminate between rabbit worms, rabbit stages and rabbit parts is no easy task. (It is instructive to compare the issue to that of the '*qua*' problems for causal theories of reference. Sterelny (cf. 1990, ch.6.))

[58] See Lewis (1974b)

[59] It should not be thought that endurantism is a more intuitively appealing doctrine either: as spelled out within the eternalist framework we are considering, it is equally, if not more, intuitively bizarre than the options just considered. See Parsons (2000). One might take the moral to be that it is the eternalist framework that is at fault: evaluating such a claim would take us far beyond the current discussion.

conception of rabbits, there seems nothing *descriptive* in ordinary conception to resolve the division question.

It seems to me that there is no reason *not* to accept division inscrutability, granted a successful resolution of 'technical' objections. Neither considerations of simplicity, nor causal constraints, promise a resolution of the issue. Equally, for the reasons just given, I think that division inscrutability is not all that counterintuitive—at least if we adopt a certain modesty about how much is (apparently) revealed in introspection about the metaphysical nature of the subject-matter of our thought and speech. Since different primitive semantic machinery is being used in the various cases, one might even suggest that it is a kind of 'framework' inscrutability mentioned at the end of Chapter §3.5; a less trivial example than the choice of analysis of relations in set-theoretic terms, perhaps, but still a matter of how we package the same ontology, ideology and meaning-facts. Appropriate objections to division inscrutability are the technical ones we have been exploring; philosophically, the case seems innocuous.

Chapter 5

Arguments for radical inscrutability

I take *positive inscrutability arguments* to have the following form:

1. SENTENTIAL CONSTRAINTS: The criterion of success for a semantic theory is that it assign the right semantic value to sentences

2. OVERGENERATION: Multiple assignments of subsentential reference generate the same semantic values

3. therefore: INSCRUTABILITY: There is no fact of the matter about what subsentential expressions refer to.

Given the versions of interpretationism currently in view, SENTENTIAL CONSTRAINTS will be sustained. This is the upshot of two theses:

1. SENTENTIAL DATA: The data-set constraining the selection of semantic theory concerns the semantic values of sentences—e.g. a set of T-sentences, or a pairing of sentences with propositions.

2. BEST=FIT: The sole criterion of success for a semantic theory is fitting this data.

We can see the Quinean 'Gavagai' arguments considered in Chapter 4 as fitting into this scheme. In the Gavagai case, I have suggested, the best course may be to accept the claim of division-inscrutability. There may be other settings—the 'innocuous' and 'illuminating' inscrutability of §3.5—where we would not wish to insist on scrutable reference in any case.

When we turn to *radical* inscrutability of reference, the claim would be that there is no fact of the matter *whatever* about which objects our words pick out: so that Paris and Sydney have as much claim as London to be the referent of "London". This is too much for most to accept. In future parts of this thesis: (1) I will be looking at whether there are good *theoretical* grounds for wanting to resist radical inscrutability (Chapters 6,7) (2) I will be looking for principled substitute for BEST=FIT that would block the move from overgeneration results to inscrutability (Chapter 8).

The aim of this chapter is to provide a battery of arguments for radical inscrutability, via outlining formal arguments for radical overgeneration. We will gain a sense of their robustness and power, and I will outline a philosophically significant (though easy) extension of the arguments. We will also see some limitations to the arguments: in particular, I describe two semantic frameworks where the arguments do not go through.

This chapter is divided into four sections. The first section details the permutation arguments for radical overgeneration popularized by Putnam (1981). The arguments are extremely simple, but powerful. I first run through the intuitive ideas, then sketch their application to (a slight generalization of) first-order predicate logic.

If we are to make a case that a natural language such as *English* is inscrutable, however, we are likely to need results in much more powerful settings. The final result we prove below is as powerful as one could reasonably wish: we work within a double-indexed general semantics of the kind described in Lewis (1970a, 1980), and show that one can permute reference schemes in arbitrary ways while leaving the pairing of sentences with semantic values (in this case, functions from indices to truth-values) fixed. Since the general semantics is multiply intensional, and contains the resources of a full type theory, this is an extremely powerful setting—corollaries will be that indexical, modal, temporal and other intensional aspects of natural language will not disrupt the permutation argument.[1]

The second (short) section describes a small extension of the argument for overgeneration. Not only can we argue that arbitrary permuted reference schemes generate the appropriate truth-conditions, but there are 'sententially equivalent' semantics that *embed reference schemes that change arbitrarily with context*. Therefore, by the same arguments as previously, interpretationism seems committed to *indexical inscrutability*: the claim that there is no fact of the matter, from one moment to the next, whether a term has retained the same referent. This result will play a central role in the discussions of Chapter 6.

The third section then looks at two settings where the permutation arguments break down. These are semantic theories formulated in terms of *structured propositions*, of the kind described by Soames (1989); and *Davidsonian truth-theoretic semantics*, (Davidson, 1967; Larson and Segal, 1995). (The latter result is perhaps surprising, since Davidson (1979) is one of few to *accept* radical inscrutability.) I shall argue that this result may be significant: it shifts potential resistance to inscrutability arguments from the relation between data and selected theory (i.e. in modifying BEST=FIT), and focuses attention rather on the kind of data that the interpretationist can provide. Only if the sentential data are 'fine-grained' (in a sense I shall spell out) do we have a chance of avoiding inscrutability. However, the leading extant accounts of such data (e.g. Lewis (1975, 1994b) Davidson (1973, 1980)) are 'coarse-grained'. Unless we provide some alternative way of identifying the data, it will be indeterminate which fine-grained description of the data is appropriate; and radical inscrutability will be re-instituted. Though the move to more fine-grained settings does point to a principled way of resisting inscrutability, hard work would be required to get it to work.

The fourth and final section looks at an alternative way of arguing for radical inscrutability of reference. Focusing in the first instance on an extensional setting, we can use the techniques used by Henkin to prove completeness and compactness theorems, to show that any consistent set of sentences can be made-true by models embedding arbitrary reference-schemes. The strategic significance of this form of argument is that, unlike the permutation arguments previously discussed, the deviant models are not characterized derivatively from an 'intended model'. This will be exploited in Chapter 8 to cause problems for the Lewis (1984; 1983a) 'eligibility' response to inscrutability. Here we describe the completeness/compactness argument, and an extension to the intensional case. In Appendix B, we provide a standard presentation of the completeness theorem for first-order predicate logic, and sketch the Henkin (1950) proof of completeness and compactness for a type theory, using general (or 'Henkin') models.

[1] Putnam (op cit) gives the permutation argument for the special case of a modal logic. Hale and Wright (1997b) give these results and also extensions to a second-order setting. The result just described subsumes these theorems. McGee (2005a) argues that the permutation arguments cannot be extended in full generality to quantification modal logic. I explain at §5.1.2 and in Appendix C how my setup differs from that that McGee presupposes, in ways that avoid his objections.

5.1 Permutation arguments

The key idea of permutation arguments is that twisted assignments of extensions (referents) to constants are compensated by equally twisted assignments of extensions to predicates. Overall, the twists 'cancel out' to deliver the usual result at the level of sentences.[2] For a toy example of this, let us use the phrase 'the image of x' to pick out *x* whenever *x* is anything other than Billy or the Taj Mahal; to pick out Billy if *x* is the Taj Mahal; and to pick out the Taj Mahal if *x* is Billy. We can describe our twisted reference scheme as follows: *N* twist-refers to *x* iff *N* standardly-refers to *y* and *x* is the image of *y*. For the similarly twisted assignment of extensions to predicates, take any atomic predicate *P*. Let *P* twist-apply to *x* iff *P* standardly applies to some *y*, such that *x* is the image of *y*. The twists cancel out—the distribution of truth-values to sentences is the same on both interpretations. For example, "Billy is running" is true iff the referent of "Billy" falls under the extension of "runs". Now, the twist-referent of "Billy" is the Taj Mahal; but the twist-extension of "runs" includes all the images of running things. Since Billy runs, and the Taj Mahal is the image of Billy, the Taj Mahal falls under the twist-extension of "runs". The sentence comes out true, just as it does under the standard interpretation. The point of the permutation arguments we give below will be to generalize this to arbitrary sentences in rich languages, and, where appropriate, to show that the permutations can preserve semantic properties of sentences beyond simple truth-value.

[2] Versions of the permutation argument are given in Jeffrey (1964); Quine (1964); Field (1975); Wallace (1977); Putnam (1978a); Davidson (1979); Putnam (1981). The most detailed presentation of the results that I know of is in the appendix to Hale and Wright (1997b).

5.1.1 Permutation arguments preserving truth-values.

The simplest form of permutation argument can be deployed against global descriptivism (cf. §1.3), where the sole constraint on successful interpretation is that it render true a certain set of sentences: "total theory". What form an interpretation should take depends on the syntactic complexity of the total theory. If the syntax is that of first-order logic, then a first-order model theory will suffice. If it includes sentential operators (e.g., "necessarily", "possibly"), then a more sophisticated semantic theory may be needed.

To deal with natural language within a model-theoretic approach, the kind of interpretation offered may need to have the richness of the intensional frameworks of Montague (1970); Lewis (1970a), discussed below. Nevertheless, in this section I will build up the permutation argument for extensional fragments of language, in order to illustrate the general theme of later constructions.

This section contains: first, a permutation argument for the quantifier-free fragment of first-order predicate logic; and second, a permutation argument for a generalization of predicate logic, where we can have 'quantifiers' other than the traditional first-order $\exists$ and $\forall$. This enables us to illustrate moves that will be used later in the extension of the argument to a general semantic setting.

Quantifier-free Predicate logic

The first framework to be considered is a language with propositional connectives, predicates (of any finite adicity) and constants, but no quantificational expressions. Complex expressions and well formed formulae are defined in the obvious way. This is the 'predicate logic' (PL) of Davis and Gillon (2004, ch.3, §3.2.1). It will suit our purposes to set up the model theory in the following 'general' way. A model will consist of the pair of a domain U and a *lexical interpretation function* $||$. A lexical interpretation function will be a function from atomic expressions (predicates, constants and propositional connectives) to appropriate extensions. As before, the appropriate extension for a constant will be an element of U, and the appropriate extension for an n-adic predicate will be a function from the nth Cartesian product of U to truth-values (i.e. a function from n-tuples of elements of U to truth-values). The appropriate extension for a n-ary propositional connective will be a function from n-tuples of truth-values to truth-values.

For m a model with lexical interpretation function $||$, we can then define the classical m-valuation of the language in the following way:

1. For Π an n-adic predicate, and c_i constants:
   $v_m(\Pi c_1 \dots c_n) = T$ iff $|\Pi|\langle |c_1|, \dots, |c_n|\rangle = T$.
2. For $*$ an n-ary connective, and α_i formulae, we have:
   $v_m(*(\alpha_1, \dots, \alpha_n)) = T$ iff $|*|(v_m(\alpha_1), \dots, v_m(\alpha_n)) = T$

Now, given a lexical interpretation function $||$, and for any permutation ϕ of U,[3] say that $||_\phi$ is the ϕ-*variant* of $||$ if it meets the following conditions:

- For Π an n-adic formula, we have:[4]
  $$|\Pi|_\phi = |\Pi| \circ \phi_n^{-1}$$
  That is, if $|\Pi| : \langle a_1, \dots a_n\rangle \mapsto T$, then $|\Pi|_\phi : \langle \phi(a_1), \dots \phi(a_n)\rangle \mapsto T$
- For c a constant, $|c|_\phi = \phi(|c|)$.

[3] i.e. a bijection from U to U.

[4] Here ϕ_n^{-1} is the obvious function from U^n to U^n induced by ϕ^{-1}. I.e. $\chi_n : \langle a_1, \dots, a_n\rangle \mapsto \langle a'_1, \dots, a'_n\rangle$ iff $\chi : a_i \mapsto a'_i$ for each i.

- For $*$ a connective, $|*|_\phi = |*|$.

Let m_ϕ be the model obtained from m by replacing its interpretation function by the ϕ-variant. It is then an easy exercise to show that v_m and v_{m_ϕ} assign the same truth-values to each formula in the language. For whereas the reference-scheme (assignment of objects to constants) has been permuted, the predicate-extensions have been permuted in compensating ways. The 'twists' of the reference scheme are 'untwisted' by the predicate extension scheme, and the overall assignment of truth-values to sentences is invariant.

More formally, one first checks the base case—the assignment of truth-values to atomic sentences. It suffices to note that:

$$\begin{array}{rcccccl|l}
v_{m_\phi}(\Pi c_1 \ldots c_m) = T & \leftrightarrow & |\Pi|_\phi & : & \langle |c_1|_\phi, \ldots, |c_m|_\phi \rangle & \mapsto T & & \text{(defn } v_x\text{)} \\
 & \leftrightarrow & |\Pi|_\phi & : & \langle \phi|c_1|, \ldots, \phi|c_m| \rangle & \mapsto T & & \text{(defn } ||_\phi\text{)} \\
 & \leftrightarrow & |\Pi|_\phi & : & \phi_n\langle |c_1|, \ldots, |c_m| \rangle & \mapsto T & & \text{(defn } \phi_n\text{)} \\
 & \leftrightarrow & |\Pi|_\phi \circ \phi_n & : & \langle |c_1|, \ldots, |c_m| \rangle & \mapsto T & & \text{(maths)} \\
 & \leftrightarrow & |\Pi| \circ \phi_n^{-1} \circ \phi_n & : & \langle |c_1|, \ldots, |c_m| \rangle & \mapsto T & & \text{(defn } |\Pi|_\phi\text{)} \\
 & \leftrightarrow & |\Pi| & : & \langle |c_1|, \ldots, |c_m| \rangle & \mapsto T & & \text{(maths)}
\end{array}$$

The last line here is just the condition for $v_m(\Pi c_1 \ldots c_m) = T$, so the equivalence for atomic formulae has been established. Since the interpretation of the logical connectives is constant, the induction steps will be trivial. This establishes the result that v_m and v_{m_ϕ} assign the same truth-value to every formula in the language.

Quantifiers

We now prove a permutation result for a quantified predicate logic (QPL). As mentioned, we will not build special clauses for the quantifiers into the definition of truth-in-a-model. Instead, we provide a general framework wherein the quantificational symbols themselves receive interpretations.

The framework for QPL has propositional connectives, quantifiers, predicates (of any finite adicity), constants and variables. Complex expressions and well formed formulas are defined in the familiar way.

A model will consist of the pair of a domain U and a lexical interpretation function $||$. As before, this gives function from atomic expressions (predicates, constants, connectives and quantifiers) to appropriate extensions. The appropriate extension for a constant will be an element of U. The appropriate extension for an n-adic predicate will be a function from the nth Cartesian product of U to truth-values (i.e. a function from n-tuples of elements of U to truth-values). The treatment of appropriate extensions for connectives differs from that in (PL), and quantifiers will find their place as a species of connective.

To set up the semantics in this general setting, we shall use the kind of 'cylindrical algebra' described by Davis and Gillon (2004, ch.3). The semantic values of sentences will be sets of *variable assignments*: functions from variables x_i to elements of U. Call the set of all variable assignments V. Intuitively, the set of variable assignments associated with an open sentence p will be all those assignments to its free variables which render the sentence true. Intuitively, a sentence will be true *simpliciter* if every variable assignment renders it true; that is, if its semantic value is V.

The values of connectives and quantifiers reflect this. The value of 'and', for example, will be intersection: for a variable assignment to render true the conjunction of two open sentences, it must render true each of them. The appropriate extension for a two-place connective is a function from $(\mathbb{P}V)^2$ to $\mathbb{P}V$. Simple quantifiers such as $\forall$ and $\exists$ then become special one-place connectives. For reference, the intended interpretations of the common logical connectives are given below.[5]

[5] Compare Davis and Gillon (2004, ch 3, Def 27).

$$\begin{array}{lcccl}
|\neg| & : & |\alpha| & \mapsto & (V - |\alpha|) \\
|\wedge| & : & |\alpha|, |\beta| & \mapsto & |\alpha| \cap |\beta| \\
|\vee| & : & |\alpha|, |\beta| & \mapsto & |\alpha| \cup |\beta| \\
|\supset| & : & |\alpha|, |\beta| & \mapsto & (V - |\alpha|) \cup |\beta| \\
|\exists x_i| & : & |\alpha| & \mapsto & \{g : \text{for some } h \in |\alpha|, g \sim_{x_i} h\} \\
|\forall x_i| & : & |\alpha| & \mapsto & \begin{cases} |\alpha| & \text{If whenever } g \in |\alpha| \text{ and } h \sim_{x_i} g \text{, we have } h \in |\alpha|. \\ \emptyset & \text{otherwise} \end{cases}
\end{array}$$

(When f, g are variable assignments, we write $f \sim_{x_i} g$ iff f and g agree everywhere except possibly on the values assigned to x_i.)

As before, given a lexical interpretation function $||$ within a model m, we move to define the valuation v_m.

1. For Π an $(n+k)$-adic predicate, and x_j variables and c_i constants:[6]
   $v_m(\Pi x_1, \ldots, x_k, c_1 \ldots c_n) =_{\text{def}} \{g \in V : \langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle \in |\Pi|\}$.

2. For $*$ an n-ary connective, and α_i formulae, we have:
   $v_m(*(\alpha_1, \ldots, \alpha_n)) = X$ iff $|*|(v_m(\alpha_1), \ldots, v_m(\alpha_n)) = X$

We call a formula α 'true' simpliciter when $v_m(\alpha) = V$.

Now we can define the ϕ-variant of $||$ in the same way as before for predicates and constants. We need to be careful with connectives. First, let the ϕ-variant g^{ϕ} of a variable assignment g be $\phi \circ g$: i.e the function that maps x_i to $\phi(a)$ iff g maps x_i to a. Then let the ϕ-variant X^{ϕ} of a set of variable assignments X be $\{g^{\phi} \mid g \in X\}$. (Notice that $V^{\phi} = V$.) Now we can set the clause for connectives. Where $*$ is an n-ary connective, let

$$|*|_{\phi} : X_1^{\phi}, \ldots, X_n^{\phi} \mapsto Y^{\phi}$$

iff

$$|*| : X_1, \ldots X_n \mapsto Y$$

(Notice that for $*$ one of the logical connectives or quantifiers, and $||$ the intended interpretation, $|*|_{\phi} = |*|$ as before. Indeed, one proposal for characterizing *logical* connectives and quantifiers is exactly this invariance under permutations.[7])

We can now prove that for any formula (open or closed), $|\alpha|_{\phi} = X^{\phi}$ iff $|\alpha| = X$. This will suffice to show that $||$ and $||_{\phi}$ agree on the truth-values of all closed sentences. A closed sentence can have only one of the two semantic values V ('truth') and $\emptyset$ ('falsity'). Since, as noted previously, $V = V^{\phi}$, and clearly $\emptyset = \emptyset^{\phi}$, the result just stated will lead to invariance of truth-value.

The base case is as follows:

$$v_{m_{\phi}}(\Pi x_1, \ldots x_k, c_1 \ldots c_m) = X^{\phi}$$

$$\begin{array}{llrcll|l}
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi|_{\phi} & : & \langle g^{\phi}(x_1), \ldots, g^{\phi}(x_k), |c_1|_{\phi}, \ldots, |c_m|_{\phi}\rangle & \mapsto T\} & (\text{defn } v_x) \\
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi|_{\phi} & : & \langle \phi[g(x_1)], \ldots, \phi[g(x_k)], \phi|c_1|, \ldots, \phi|c_m|\rangle & \mapsto T\} & (\text{defn } g^{\phi}) \\
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi|_{\phi} & : & \phi_n\langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle & \mapsto T\} & (\text{defn } \phi_n) \\
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi|_{\phi} \circ \phi_n & : & \langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle & \mapsto T\} & (\text{logic}) \\
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi| \circ \phi_n^{-1} \circ \phi_n & : & \langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle & \mapsto T\} & (\text{defn } ||_{\phi}) \\
\leftrightarrow & X^{\phi} = \{g^{\phi} \mid & |\Pi| & : & \langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle & \mapsto T\} & (\text{defn logic}) \\
\leftrightarrow & X = \{g \mid & |\Pi| & : & \langle g(x_1), \ldots, g(x_k), |c_1|, \ldots, |c_m|\rangle & \mapsto T\} & (\text{defn } X^{\phi})
\end{array}$$

$$\leftrightarrow v_m(\Pi x_1, \ldots x_k, c_1 \ldots c_m) = X \qquad (\text{defn } v_x)$$

[6] Note that we may need to reorder to get to this case: allowing for this will be routine.

[7] The proposal is due to Tarksi. See MacFarlane (2000, ch 5.) for extensive and fascinating discussion.

This gives us the base case.

For the induction step, we need to ensure that

$$|*|(|\alpha_1|,\ldots,|\alpha_n|) = X \Longleftrightarrow |*|_\phi(|\alpha_1|_\phi,\ldots,|\alpha_n|_\phi) = X^\phi \tag{†}$$

The induction hypothesis is:

$$|\alpha_i| = X \Longleftrightarrow |\alpha_i|_\phi = X^\phi \tag{††}$$

We also know, by construction of $||_\phi$, that:

$$|*|(X_1,\ldots,X_n) = X \Longleftrightarrow |*|_\phi(X_1^\phi,\ldots X_n^\phi) = X^\phi \tag{‡}$$

by the induction hypothesis, ††, the right hand side of (†) may be rewritten $|*|_\phi(X_1^\phi,\ldots,X_n^\phi) = X^\phi$, where $X_i = |\alpha_i|$. This, in the presence of (‡), is equivalent to $|*|(X_1,\ldots,X_n) = X$; and we have our result.

Recapitulation

We have shown how the permutation arguments go through in a simple way for a quantifier-free language; and how these arguments can be generalized to (an extension of) first order predicate logic. This gives us an argument for radical overgeneration, and thus for radical inscrutability of reference, within the setting of global descriptivism (§1.3).

The settings so far covered are somewhat limited. Rather than develop analogous results for higher order, modal, tense, and indexical logic, and combinations thereof, we shall deal with them all at once. This will be achieved by proving a more general result, which shows that a more general property of sentences can be held invariant under permutations of the reference scheme. This general result will give a proof of radical overgeneration, not just for the global descriptivist setting where invariance of *truth-values* of sentences suffices; but for the richer interpretationist settings where we need to secure invariance of *truth-conditions*.

We shall ultimately provide a permutation result for a double-indexed general semantics. This is a higher order, multiply intensional setting which is can represent the indexical character of a language. We shall see that a permutation of the reference (extension) of singular terms (terms of category N) can leave invariant the intension assigned to sentences (symbols of category S). To say that the intension of a (closed) sentence is invariant, is to say that it takes the same truth-value relative to each index—a tuple of possible world, time, variable assignment etc. *A fortiori*, it will take the same truth-value at the actual world, present time, given any variable assignment. As a corollary, it leaves the truth-values invariant.

5.1.2 Permutation arguments preserving truth-conditions

In some interpretationist settings, one's data consist of a pairing of sentences with coarse-grained propositions or intensions (cf. the Lewisian interpretationism described in §1.3). The target is to preserve, not only the truth-values of certain sentences, but the *truth-conditions* of sentences. In the current setting we shall think of truth-conditions as given by sets of possible worlds, or characteristic functions of such sets: this tells us *the worlds in which the sentence is true*. (In §5.3 we shall consider two more fine grained notions of 'truth-conditions' one might use, which significantly alter the standing of the permutation argument).

We will show how arbitrary permutations of the extension of singular terms can be embedded within an overall interpretation which leaves the truth-conditions assigned to sentences invariant. In fact, the setting is more general. The semantics will assign to sentences functions from indices to truth-values, where the indices specify not only a possible world, but also a time, place, and other factors. As explained at §5.1.2, if we can show that the permuted variants of an interpretation leave the semantic value of sentences invariant, we shall have what we need.

We first prove the result for the general semantic setting of Lewis (1970a) (what Cresswell (1973) calls a 'pure categorial language').[8] We first set out the framework itself, and then state and prove a permutation theorem for the setting. We shall then describe how the framework is altered to take into account indexicality, and extend the permutation argument to that 'double-indexed' setting.

The framework (syntax): a pure categorial language

The framework in which we operate initially is a *general semantics*, in the sense of Lewis (1970a). The atomic expressions of the language will be given by a *lexicon*, and there will be phrase-structural rules saying how admissible compounds can be built up from elements of the lexicon. The lexicon and the phrase-structural rules give the analogue of a definition of well-formedness within a formal language.

The framework that Lewis recommends is drawn from Ajdukiewicz's (1935) categorial grammar, and is what Cresswell (1973) calls a 'pure categorial language' (essentially, this is a form of type theory, without any 'syncategorematic' terms such as λ-operators). In the lexicon, atomic expressions will be associated with *categories*, and then the well-formedness of compounds will be determined by whether the categories of the components dovetail. For example, the category of the lexical item "Susan" might be that of 'names' (N), and the category of the lexical item "runs" might be 'intransitive verb' (S/N)—the notation reflecting the fact they when combined with a category N expression it produces a category S expression (a sentence).

The categories themselves are characterized recursively. N and S are 'basic categories'[9] Given categories $c, c_1, \ldots, c_n$, we have a *derived* category $c/c_1, \ldots, c_n$. In the above, 'runs' is in the derived category S/N. Writing "+" for concatenation, the general rule of well-formedness is that for any expressions $e, e_1, \ldots e_n$, we have $e + e_1 + \ldots + e_n$ well-formed iff e is of a derived category $C/C_1, \ldots, C_n$ and each e_i is of category C_i. "runs(Susan)" is well-formed because it is of the form $(S/N) + N$, and so fits our template.[10]

Notice that the complexity of the category of an expression and its syntactic complexity cross-cut. Basic lexical items—the atomic expressions—will typically have quite complex categories: for example, the atomic adverb 'fast' will have category $(S/N)/(S/N)$ (it yields a verb phrase when a verb phrase is input). Conversely, expressions of the basic category S (sentences) will be syntactically complex.

[8] In Appendix C we extend this to a version of Cresswell's λ-categorial languages.

[9] In fact, it is controversial whether ordinary proper names are in category N.

[10] I keep the bracketing for notational convenience, though it plays no part of the official definition. It would be a trivial matter to alter the definition of well-formedness to allow for parentheses, but it would be a distraction.

The result of such concatenation won't look much like English. A very basic example is that 'Jill loves Jane' would be represented by the word-order '(loves(Jane))Jill' within the system just developed. Additional work is required to explain how such analyses relate to natural language. What we need here is some kind of mapping from our 'disambiguated' formal language to ordinary sentences of English. This 'ambiguating relation' may take one of a number of forms, depending on what other resources from linguistics one wishes to bring to bear. Lewis (1970a, p.204) appeals to a "transformational component" of the grammar; Dowty (1979) suggests various ways to integrate the setting with generative semantics.[11]

The framework (semantics): single-indexed general semantics

What we have so far seen is a treatment of the syntax of a formal language. We now describe the Lewisian "general semantics" which will associate each well-formed expression with a semantic value. The kind of semantic value for a given expression will be what Lewis calls an "appropriate intension"—which intensions are appropriate being determined recursively by the category of the given expression. The base cases are handled individually. For sentences (expressions of category S) an appropriate intension will be a function from indices to truth-values. For expressions of category N, it will be a function from indices to objects.[12] Appropriate intensions of derived categories are then functions between intensions appropriate to the categories from which they are derived. For example, expressions in category S/N will be assigned functions from intensions appropriate to category N expressions to intensions appropriate to category S.

We can capture this by characterizing intensions in terms of their *type*. Let indices be of type i, objects type o and truth-values type t.[13] Let the type $\langle a,b\rangle$ include all functions from elements of type b to elements of type a. We can then formulate the above by saying that an appropriate intension for category S is a function of type $\langle t,i\rangle$; an appropriate intension for the category N is of type $\langle o,i\rangle$; and in general the appropriate intension for category c/d, where c is of type α and d is of type β, will be of type $\langle\alpha,\beta\rangle$.[14][15]

A *lexical interpretation* $||$ will be an assignment of semantic values to basic parts of the language—the elements of the lexicon (e.g. names, adjectives, verbs, adverbs etc). We now have the idea of a lexical interpretation function for a categorial language, and a characterization of the *type* of semantic

[11] Montague used a richer set of rules *within* the general semantics itself to make the formal language closer to English syntax. Cf. (cf. Partee, 1996, §3), Dowty (1979, ch.1) for discussion.

[12] Lewis (1970a) includes category C (common nouns) as another basic category, having as appropriate intension a function from indices to sets of objects. For the sake of simplicity, I ignore this here.

[13] Note that there is a single 'type' containing all objects—when we define intensional operators over this framework, we will end up with a 'single domain' semantics. Plausibly, we want to be able to say that there are things which do exist, which might not exist had some other situation been actual. The solution is to pair each world-index w with a partition of the type of objects into two: those that exist at w, and those that do not. We then interpret the 'ordinary' loaded existential quantifier relative to w be restricted to those objects that exist at w. Cf. §5.1.2, below.

[14] This is formulated for derived categories with two elements only. For 'binary branching' languages we need no more. For the more general case, we need some additional notation to mark functions from sequences of types to types. (McCawley, 1980, ch.13)

[15] The compositional intensions which Lewis favours differ from the 'Carnapian intensions' that form the basic framework for Montague (1970). To characterize a Carnapian intension, one first defines the notion of an appropriate *extension* in a way paralleling the recursive formulation above. The extension for a name will be an individual of type o; the extension for a sentence will be a truth-value of type t, and in general an appropriate extension for an expression of category c/d will be of type $\langle\alpha,\beta\rangle$, where α,β are the types appropriate to c,d respectively. An appropriate *Carnapian intension* for an expression of category Q/R will be a function from indices to appropriate extensions, i.e. of type $\langle i,\langle\alpha,\beta\rangle\rangle$ where $\langle\alpha,\beta\rangle$ is the appropriate extension for category Q/R. The deficits of a framework based on Carnapian intensions—centrally, its inability to deal with 'intensional' expressions such as the predicate 'is rising'—are discussed in Lewis (1970a) and are addressed by Montague by assigning certain expressions semantic values which are not of the Carnapian kind. See Thomason (1974b).

value appropriate to complex expressions. We need to 'extend' the lexical interpretation function to give an overall assignment of semantic values to arbitrary well-formed expressions in the language.

We let a model m for the language include a domain U, a set of indices I, and a lexical interpretation function $||$. Then one could let a model induce a valuation v_m for arbitrary well formed expressions:

1. For e in the lexicon, $v_m(e) = |e|$

2. For $e' = e(e_1, \dots e_n)$, $v_m(e') = v_m(e)[v_m(e_1), \dots, v_m(e_n)]$

In what follows, I will choose a slightly different, though equivalent, setting. I will speak of an arbitrary *valuation* $||$ as any function taking expressions to intensions appropriate to their category. Such valuations may very well be totally crazy: what we typically do is to restrict attention to the well-behaved ones. I say that $||$ *embeds* a lexical interpretation scheme if that scheme coincides with the restriction of $||$ to lexical elements. I call a valuation $||$ *compositional* if the following general relation holds:

$$|e(e_1, \dots, e_n)| = |e|[|e_1|, \dots, |e_n|]$$

We then let models contain compositional valuations, rather than lexical interpretation schemes.[16]

This finishes the description of the syntax and semantics of our 'single indexed pure categorial language'. To summarize:

1. Our language consists of infinitely many categories of expressions, described via a categorial grammar. S and N are basic categories. If C and $C_1, \dots, C_n$ are categories (basic or derived), let $C/C_1 \dots C_n$ be a derived category.

2. We have a set of *indices*, I and a domain of *objects* O, a set of *truth-values* $\{T, F\}$, and a valuation $||$.

3. The valuation $||$ assigns intensions to each expression appropriate to its category. *Appropriate intensions* for each category are functions whose type reflects the build up of the category, in the fashion described above.

4. Call a valuation *compositional* iff the semantic projection rule of function-application is met. Suppose we have an expression e of category $C/C_1 \dots C_n$, and expressions e_i of categories C_i (so that the complex expression $e(e_1 \dots e_n)$ will be of category C. Then it must be that:

$$|e(e_1 \dots e_n)| = |e|(|e_1|, \dots, |e_n|)$$

The permutation argument within single-indexed general semantics

Let us introduce two further notions. First, call two compositional interpretations *sententially equivalent* if they assign the same intensions to everything of category S. Second, say that a valuation embeds the reference scheme r (relative to index i) iff whenever e is of category N, $|e|(i) = r(e)$. The aim now is to show that we can always construct *sententially equivalent* compositional valuations which embed arbitrarily permuted reference schemes.

[16] The two settings are equivalent. Given a model that supplies a lexical interpretation $||$, the corresponding compositional valuation is the valuation v_m determined by $||$. Given a model that supplies a compositional valuation, the restriction of that valuation to atomic expressions is the corresponding lexical interpretation. A minor advantage of this setting is that even semantic projection rules are no longer required within the semantic theory, minimizing 'syncategorematic' elements that are unmotivated in a metasemantic setting.

The theorem we will prove shows that any coherent[17] assignment of semantic values (i.e. appropriate intensions) to sentences can be made compatible with arbitrarily permuted reference schemes. We do this by defining the notion of a ϕ-variant of a valuation, where ϕ is a permutation of the domain of objects. If $||$ is a valuation embedding at index i the reference scheme r that, e.g. assigns Susan to "Susan"; then the ϕ-variant $||_\phi$ of $||$ will embed at c a reference-scheme r_ϕ that assigns ϕ(Susan) to "Susan". Rather than setting up a permuted lexical interpretation scheme and proving by induction that the valuations determined by the original and ϕ-variants appropriately match, I shall define directly the ϕ-variant valuations, in a way that makes obvious the fact that they are sententially equivalent to the valuation from which one starts, and that they embed an appropriate permuted reference scheme. The challenge will then be to show that the definition is legitimate: i.e. that the permuted valuation, so-defined, is compositional. We must show that the semantic values that it assigns to complex expressions arise in the appropriate way from the semantic values it assigns to the lexical basis.

The construction of the ϕ-variant valuations is as follows:

- Let ϕ be an arbitrary permutation of O. Define recursively the ϕ-image of an intension as follows:[18]
  - If f is an appropriate intension for S, then $f^\phi = f$
  - If f is an appropriate intension for N, then $f^\phi = \phi \circ f$
  - If f is an appropriate intension for $C/C_1 \ldots C_n$,
    then $f^\phi : r^\phi \mapsto g^\phi$ iff $f : r \mapsto s$.
- Given a valuation $||$, let the ϕ-permuted valuation $||_\phi$ assign to e its ϕ-image; i.e. for each e, $|e|_\phi := f^\phi$ where $f = |e|$.

Notice that, by construction, a ϕ-variant of a valuation v will embed a reference scheme that is permuted by ϕ. Moreover, a valuation and its ϕ-variant will be sententially equivalent, again by construction. What remains to be shown is that the ϕ-variant is a legitimate valuation of the language—that is, that it is compositional in the sense delimited above. That isn't hard:

Theorem 1. *For any permutation ϕ, $||_\phi$ is compositional if $||$ is.*

Proof. Suppose $||$ is compositional. Take an expression e of category $C/C_1 \ldots C_n$, and expressions $e_1 \ldots e_n$ of categories $C_1 \ldots C_n$ respectively. The complex expression $e(e_1 \ldots e_n)$ will be of category C). By compositionality of $||$, we have:

$$|e(e_1 \ldots e_n)| = |e|(|e_1|, \ldots, |e_n|)$$

We need to show that:

$$|e(e_1 \ldots e_n)|_\phi = |e|_\phi(|e_1|_\phi, \ldots, |e_n|_\phi)$$

We show this for the case $n = 1$, the other cases being trivial extensions. Write $e_1 = d$. From the compositionality of $||$ we have: $|e(d)| = |e|(|d|)$; i.e. $|e| : |d| \mapsto |e(d)|$ By the definition of the ϕ-image this gives us: $|e|^\phi : |d|^\phi \mapsto |e(d)|^\phi$. Given the way that $||_\phi$ is defined, we can rewrite this as: $|e|_\phi : |d|_\phi \mapsto |e(d)|_\phi$, i.e.:

$$|e(d)|_\phi = |e|_\phi(|d|_\phi)$$

so compositionality is secured. □

[17] Call an assignment of semantic values to sentences *coherent* if there is at least one compositional valuation that embeds that assignment.

[18] If, following Lewis (1970a), we included an extra basic category C ("common nouns") mapping indices to sets of objects, then the permuted variant of $|e|$ for e in category C should be $|e| \circ f^1$.

The permutation argument within double-indexed general semantics

Indexical expressions are commonly thought to require the semantic values we have been discussing to be changeable: an expression may express different such values when uttered in different contexts. One option within a general semantics is to include context (represented as a 'centred world') as one index within the above account.[19] Within that setting, our proof runs just as before.

Alternatively, one can separate world, time, delineations etc. from context, to "double index" the general semantics.[20] Syntactically, everything is unchanged. Within the semantics, however, expressions are assigned *characters* rather than intensions, where a character is a function that maps contexts to Lewisian compositional intensions. We need the notion of an 'appropriate' character for a given category of expression. Therefore, say that a character is of type τ iff it maps contexts to intensions of type τ; and wherever the appropriate intension for category c was type τ, appropriate characters will be those of the same type.

Formally, characters resemble 'Carnapian intensions' rather than the 'compositional intensions' of Lewis' general semantics. Consider, for example, an intransitive verb (category S/N). In Lewis' general semantics, its semantic value was the function from semantic values of type N to semantic values of type S. On the double-indexed proposal, by contrast, its semantic value is a function from contexts to intensions; i.e. the context index is treated in a 'Carnapian' way.[21] The upshot of this is that the semantic projection rule cannot simply be function-application, as within the original Lewis treatment. We can still have a single rule appropriate to arbitrary concatenations, however. Consider an expression $a(b)$, where b is of category c_1 and a is of category c_2/c_1. If a has character A and b has character B, then the character of $a(b)$ maps context α to the intension that results from "compositionally" applying $A(\alpha)$ to $B(\alpha)$.

We need to spell out what the permuted interpretation function is for this new setting. This is defined in the obvious way. If $|e|$ is σ, where $\sigma : c \mapsto g$, then $|e|_\phi$ is σ^ϕ, where $\sigma^\phi : c \mapsto g^\phi$, where g^ϕ is the ϕ-image of the intension g. Notice that since $g^\phi = g$ when e is a sentence we have that $|e|_\phi = |e|$ for e a sentence. The permuted interpretation not only holds the intensions of sentences invariant, it holds the characters invariant also.

Derivative invariance results

We have seen that we can leave the semantic values of sentences invariant under permutations of the reference-scheme in a rich class of settings. The ultimate aim of such results is to argue for radical inscrutability, by saying that all such 'sententially equivalent' semantic theories are equally good at fitting the data which the interpretationist provides. Such data may not take the form of a pairing of sentences with something as rich as the compositional intensions or characters that we have just been looking. It might rather involve a pairing between a sentences and propositions—the set of circumstances in which the sentence is to be true.

In such a setting, what we need for our purpose is that semantic theories that differ by containing permuted reference-schemes can assign to sentences the same truth-conditions (function from possible worlds to truth-values). This follows immediately from the above result once we note that the truth-conditions of a sentence are determined by the semantic value of that sentence in the rich settings just

[19] See Lewis (1970a, appendix).

[20] Consider, for example, what happens if context changes through the utterance of a sentence say (cf. Lewis, 1979c, p.241). So long as we can assume that there's a unique context associated with each lexical element, context will determine an intension for each such element, which then collectively determine an intension for the whole. The story would be far more complex if we proceed in terms of functions from context-index pairs to truth-values.

[21] Cf. Appendix C

canvassed. If semantic theories are sententially equivalent, then they will assign to sentences the same truth-conditions.

To illustrate this, let us show how to construct two senses of 'truth-conditions' for a sentence S. These are what are known within 'two-dimensional modal logic' as the C-intension and the A-intension of a sentence.[22]

For w a centred world (i.e. a context), define the matching-relation $m(x,y)$ to hold between world w and string of indices i iff the world, time etc. components of i match those given by w. First, we will define the C-intension of a sentence s at a context c. This is a set of centred worlds given by:

$$w \in C(s) \Longleftrightarrow \forall i(m(w,i) \rightarrow (|s|(c) : i \mapsto T))$$

Second, we will define the A-intension of a sentence. This is the set of centred worlds given by:

$$w \in A(s) \Longleftrightarrow \forall i(m(w,i) \rightarrow (|s|(w) : i \mapsto T))$$

To see the difference between these two constructs, consider the sentence "he is Beckham", uttered while pointing at the famous footballer. The C-intension is the set of worlds where the person indicated (i.e. Beckham himself) is identical to Beckham. On standard assumptions, then, this is the necessary proposition—the set of all possible worlds.[23] The A-intension is the set of worlds where *the person indicated by the pointing* is Beckham. It is false, therefore, with respect to possible worlds where Beckham stayed at home and sent out his double to make the public appearance. Notice that the A-intension is contingent where the C-intension is necessary. It is arguable, in the light of these distinctions, that the data for a semantic theory might be given in the form of a pairing of sentences with A-intensions.[24] It is reassuring for the inscrutabilist, then, that all these sentential constructions are invariant under the permuted valuations. This can be easily checked by noting that the only appeal to the valuation $||$ in the above definitions is in application to *sentences*, where, as we have seen, $||$ and $||_\phi$ coincide.

Inscrutability of existence.

I note one last inscrutability result. We have set up our general semantics in terms of a single domain of objects (i.e. those entities of type o) from which we draw the semantic values. However, objects go in and out of existence over time; some may not exist in other possible situations, while other 'mere possibilia' do not exist in the actual world. This suggests that relative to each possible world, time, etc. we identify a subdomain of the entities of type o—those that exist in that world or time. 'Ontologically loaded' quantification would be quantification indexically restricted to those objects existing at the world, time etc. of the context.

I see no reason to suppose that we can only refer to what exists at the present time and the actual world. 'Beethoven' refers to the composer, I take it, even though he is no longer around. Salmon (1998, p.286-7) has suggested that we can construct uniquely identifying descriptions of mere possibilia and that by exploiting these we can directly refer to such objects. Similarly, objects can fall under predicates relative to times (and perhaps worlds) where they do not exit. Beethoven currently satisfies 'is

[22] For the terminology, see Jackson (1998). For alternative labels ('secondary intension' and 'primary intension' respectively), see Chalmers (1996). The C-intensions also correspond to what Stalnaker (1978) calls the 'horizontal proposition' associated with a sentence. Note that A-intensions/primary-intensions are not the diagonal propositions of Stalnaker (1978). Stalnaker's diagonal propositions are not semantic constructs, but are defined pragmatically in terms of agents' common knowledge of the semantic content of their words.

[23] I set aside concerns about worlds where the individual does not exist.

[24] See Rayo (2004) for discussion.

famous' (Cresswell, 2004). More controversially, we might think that any object whatever satisfies 'is self-identical' relative to any world or time, whether or not it exists there.[25]

Now the permutation arguments given above are indifferent to whether the image of an entity under a permutation exists in the same times and worlds as the original. One might, for example, choose a permutation that maps all objects existing in the actual world at the present time, to mere possibilia. Our result goes through just the same. 'Ontologically loaded' quantification will now range over the ϕ-images of currently existing objects; and under this interpretation, all our singular terms refer to objects not in the present time, or in the actual world. One upshot of the inscrutability arguments, therefore, is that it is inscrutable whether or not we are referring to present or even to actual objects.[26]

5.1.3 Conclusion

As promised, we have seen how to embed arbitrarily permuted reference schemes within sententially equivalent general semantic accounts of language.[27] This not only gives the overgeneration result for those interpretationisms whose data consist of a pairing of sentences with propositions, it also discharges the obligation incurred in the set-of-sentences setting, to show how permuted reference schemes can engender equivalent distributions of truth-values of sentences in expressively powerful languages. As well as the semantic values of permuted valuations matching, various other derivative sentential semantic properties match. Within the kind of framework we have been considering, the case for radical overgeneration, and so for radical inscrutability within an interpretationism that accepts BEST=FIT looks watertight.

[25] Any intuitive resistance to these claims comes, I think, from doubts about whether such entities really exist *to be* referred to. However, if we are using a possible world semantics, such entities are already within the intensions that we assign to the expressions in our language—hence we are already committed to their being able to stand in relevantly similar semantic relations. Why should being in the transitive closure of the *intension* of an expression be less ontologically worrying than being in the transitive closure of its *extension*? The appropriate response in each case is to give a satisfying story about possibilia talk (cf. p.13, above).

[26] I was led to thinking about this result by reading McGee (2005a). McGee imposes the restriction that relative to w, terms can only refer to entities in w; and uses this as an objection to permutation arguments. As noted above, I find the restriction unmotivated within a setting which allows the permutation arguments to get started.

[27] We have formulated the argument in terms of permuted reference schemes—i.e. the objects denoted, in a context, by expressions of category N. Within the general semantic framework, it is arguable that there are *no* natural language expressions of category N. Names, with noun-phrases in general, would fall within the derived category $S/(S/N)$ (Cf. McCawley, 1980, ch.13) Our results can be re-stated to finesse such issues. Each name n has a semantic value that we can think of, intuitively, as 'being one of the properties of n'. The effect of the permuted valuation schemes will be to assign it instead 'being one of the properties of $\phi(n)$'. The systematic correspondence between such values and the objects n, $\phi(n)$ respectively allows us to reconstruct the notion of 'name-reference', in terms of which the inscrutability of reference can be formulated. Compare the treatment of extensional transitive verbs in McCawley (1980, ch.13).

5.2 Indexical-inscrutability: a cut-and-shunt argument

We have examined arguments focused on what might loosely by termed "synchronic" inscrutability of reference.[28] We now prove a "diachronic" version. An expression such as "London" will, on the standard interpretation, designate London no matter what the context is.[29] Likewise, the deviant interpretations assign to that term a single object—Sydney, say—at all contexts. Because the reference of a term is at all times the image under the permutation of its original reference, the pattern of indexical dependence of reference will be the same on each of the permuted interpretation schemes.

We now have the resources to extend the arguments so that they imply a kind of 'indexical inscrutability'. The *radical indexical inscrutability* thesis is that, if c and c' are distinct contexts, then there is no fact of the matter whether e designates the same object at c as at c'. To get this result, we 'cut-and-shunt' some of the permuted interpretations constructed above. The intuitive case is straightforward. By the results of §5.1, we have two valuations that always agree with each other on the content assigned to whole sentences, no matter what context is chosen. If we construct a third valuation, which matches valuation-1 at some contexts, and valuation-2 at the remainder, it will also have this property—all three will agree on the content of sentences at each context. Our first valuation might have it that 'London' denotes London at all contexts, whereas our second valuation has it that it denotes Paris at all contexts. The constructed valuation will agree with the first on some contexts, with the second on others—so it depicts 'London' as indexical. Yet it is sententially equivalent to the originals, and so under the current assumptions, the interpretationist has no grounds for ruling it out. What we do below is sketch how this intuitive idea could be formalized within the double-indexed semantic setting given above.

Take the candidate double-indexed valuation, $||^2$. Relative to a given context c, the designated valuation determines what would be a valuation within the *single*-indexed semantics, i.e. an assignment of compositional intensions to expressions:

$$|e|^{\langle 1,c\rangle} := |e|^2(c)$$

We have seen how to construct sententially equivalent ϕ-variant interpretations $||_\phi$ of a arbitrary single-indexed valuation $||$ in the proof of the permutation result for the single-indexed case, and we can apply these techniques here, yielding the ϕ-variants $||_\phi^{\langle 1,c\rangle}$.

Now assign permutations ϕ_i to contexts via a function $\mathbb{P}$. Relative to $\mathbb{P}$, we construct a new double-indexed valuation, which we call the $\mathbb{P}$-variant of $||^2$ and denote $||^2_\mathbb{P}$.

Given an expression e and a context c, we let $||^2_\mathbb{P}$ assign to an expression e the intension given by $||_\phi^{\langle 1,c\rangle}$ at e, where $\phi = \mathbb{P}(c)$, i.e.:

$$|e|^2_\mathbb{P}(c) = f \Longleftrightarrow |e|^{\langle 1,c\rangle}_{\mathbb{P}(c)} = f$$

We now argue that $||^2_\mathbb{P}$ is sententially equivalent to $||^2$. At any given context c, the intension assigned to a term is that assigned to it on some $||_\phi^{\langle 1,c\rangle}$. However, the intension assigned by $||_\phi^{\langle 1,c\rangle}$ to a sentence is exactly that assigned to it by $||^{\langle 1,c\rangle}$, by our previous results. By construction $||^{\langle 1,c\rangle} = ||^2(c)$, so for all contexts c, $||^2(c) = ||^2_\mathbb{P}(c)$. QED.

By choosing $\mathbb{P}$ appropriately, we get radical indexical overgeneration. Consider "London", which we shall suppose to be of category N. Given contexts c and c' pick $\mathbb{P}$ where $\mathbb{P}(c)$(London)=London and $\mathbb{P}(c')$(London)=Paris. Then on the $\mathbb{P}$-variant interpretation, "London" will pick out London as uttered

[28] The terminology would be strictly appropriate if the only variable aspect of context were temporal. Clearly there are many non-temporal ways in which contexts can vary.

[29] This follows from the orthodox treatment of terms as having 'constant character', following Kaplan (1989b). Some, such as Jackson (1998), claim that the names are indexical. They maintain that, at the least, they can vary depending on what world they are uttered in.

at c and Paris as uttered at c'. Since the variant interpretation is sententially equivalent to the original interpretation, given BEST=FIT, it is not determinately incorrect.

Equally, by choosing the contexts and permutations carefully, we could take an indexical expression and give it a constant reference. One could choose $\mathbb{P}$ so that the permutation ϕ appropriate to a given context c maps the speaker of that context to London. Relative to that valuation, 'I' is a non-indexical expression referring to London irrespective of context. Again, by the argument above, such interpretations are not determinately excluded.

Indexical inscrutability, though a straightforward extension of the basic permutation argument for radical inscrutability, may be thought much more problematic: I spell this out in Chapter 7.[30]

[30] This argument will generalize to any setting where semantic values take a "Carnapian" form (i.e. they are functions from indices to some compositional value). One would take $||^2$ as a valuation appropriate to a Carnapian intensional type theory, and $||^1$ as an assignment of appropriate extensions to expressions, and appealing to a permutation result for the extensional setting (which can be taken to be a limiting case of the permutation result for the single-indexed general semantics above, where the set of indices is null). The above would give us a 'cut-and-shunt' argument for radical *intentional* inscrutability: we can embed an arbitrary assignment of intensions to singular terms. This is because we could choose different permuted assignments of extensions at different worlds. Within our setting, where we have compositional rather than Carnapian intensions, radical intentional inscrutability cannot be established by these techniques. See Appendix C for discussion.

5.3 Two settings where the argument breaks down

In the preceding sections, we have seen arguments for radical inscrutability within interpretationisms based on model theoretic semantics, whether based on a pairing of sentences with truth-values or with truth-conditions. We now turn to two settings less congenial to the arguments. These are, first, a semantics framed in terms of *structured propositions*; second, Davidson-style truth-theoretic semantics. In each case, we will find that though the arguments for overgeneration break down, the problems are relocated to a previous stage in the interpretationist project; so that the threat of inscrutability is not avoided.

5.3.1 Structured propositions

One might think that the propositions expressed by sentences must be more fine-grained than those that the general semantics just described provides for. Soames (1989) develops an account of semantic content in terms of *structured propositions*. Structured propositions, in the relevant sense, are set-theoretic constructs, language-like in structure, containing objects and properties as constituents.

The framework: structured propositions as truth-bearers

The semantic account is two-fold. First, one recursively specifies which structured propositions are associated with which sentences, by appeal to the semantic values of the atomic parts of the sentence and their mode of composition. Then, one defines what it is for a proposition to be true, and to have certain truth-conditions.

These semantic values may be appropriate 'compositional' intensions of the form just canvassed, or they could be something that determines these extensions. From a technical point of view, we need the entities to carry enough information to be able to extract the kind of semantic values we have been working with above.

For example, one could postulate *sui generis* abundant 'properties' as the semantic value of predicates, whose pattern of instantiation in possible worlds corresponds to a certain function from worlds to sets of objects (what is sometimes known as a Carnapian intension for the predicate). Given an account of cross-world identity (even a trivial one!) picking out an object would determine a name-intension—selecting that object in each world where it exists. We could then have the structured propositions contain these objects and properties, rather than the set-theoretical semantic values we have been hitherto considering. We might call these 'Russellian propositions'.

An alternative is just to put compositional intensions directly into structured propositions: these would then become what Lewis (1970a) calls 'structured intensions'.[31] Various intermediate positions are of course possible. Without further constraints provided by metaphysics, or by a theoretical role for propositions outside pure semantics, there seems little to choose between these formulations. I'll address the Russellian version, but nothing should hang on this choice.

[31] Structured intensions are really the same thing as the sentences of the 'Lagadonian language' of §2.4, above. In neither case do we need to assume that the constructs are intrinsically representational; what we appeal to is a rather convenient way of defining parochial semantic properties. Of course, if one takes propositions/Lagadonian sentences to be the real truth bearers, one might object to the term 'parochial', thinking of these as the only legitimate notions of 'truth' etc. If so, the metasemantic challenge for the semantic property of truth may be easy; the difficult thing is the semantic property of *expressing the proposition p*. Given a general semantics for a language, and so an assignment of semantic values to the atomic parts of language, it is easy enough to define this relation, so our interpretationist project can be adapted to this reformulation of the problem of intensionality anyway.

The standing of permutation results and the additional premiss.

Reformulating the "semantic values" of sentences as Russellian propositions *embedding* the objects referred to by the terms within that sentence has a dramatic effect on the permutation argument. The Russellian proposition for "Susan runs" on a standard interpretation might be:

$\langle$Susan, running$\rangle$

Writing (running)$^{\phi^{-1}}$ for the property that applies to all those things that are ϕ-images of runners, the Russellian proposition of the sentence on a ϕ-variant interpretation would be

$\langle\phi$(Susan), (running)$^{\phi^{-1}}\rangle$

Clearly, the proposition paired with the sentence has altered: within the framework of structured propositions, the permuted reference schemes do impact on the semantic values of sentences.

One might wish to dispute this point. One might claim that we can make the valuation $||$ and its permuted variant $||_\phi$ deliver the same result, by putting appropriate twists into the clauses that say how the structured propositions are determined by the semantic values of the parts of a sentence. For example, we have been assuming:

If $|c| = x$, and $|\Pi| = P$, then $|\Pi c| = \langle x, P\rangle$

Why must this hold? Why isn't an alternative the following:

If $|c| = x$, and $|\Pi| = P$, then $|\Pi c| = \langle \phi^{-1}(x), (P)^{\phi}\rangle$

If we can include the clauses of the former kind along with the initial valuation, and clauses of the latter kind with its ϕ-variant, then the structured propositions associated with sentences match.

This argument strikes me as a cheat. In the case at hand, we seem free to stipulate the form that the compositional clauses are to take—they are part of the semantic framework, the terms in which we formulate an interpretation, rather than being a part of the interpretation itself.[32]

Have we found the key flaw in arguments for radical inscrutability? Is avoiding such arguments as easy as choosing a structured-propositions semantic setting rather than the general semantics previously discussed? It must be admitted that there is a programme of interpretationism whereby this could be exploited. The crucial move would be to argue that, on an appropriately non-semantic basis, one could extract a pairing of sentences with Russellian propositions that will constrain re-interpretation in just the way envisaged. Herein lies the catch: the structured-proposition framework can resist permutation moves just because it postulates a far richer data-set than that appealed to by its rivals. In principle, such a theorist could appeal to Lewis' convention-based account: an appeals to a convention to utter "Susan runs" iff one believes the *structured* content $\langle$Susan, running$\rangle$. In order for this to fit into an overall reductive account of intentionality, however, one would have to say what it is for someone to believe structured propositions. "Head-first" theorists such as Lewis and Stalnaker typically assign mental content in a less finely grained way—the kind of counterfactual and decision-theoretic resources they appeal to will not

[32] Not all clauses that are part of the theory rather than settled by the interpretation function could be defended in this way: often one finds specific axioms governing the behaviour of compounds such as $A \wedge B$, rather than this being determined by what the interpretation function assigns to $\wedge$. In such cases, we are entitled to extend the interpretation function in order to eliminate such parochial axioms. The process has to end somewhere and there is no motive in the case at hand for allowing re-interpretation.

Where to stop is a matter of controversy: consider second order logics where much turns on whether the concatenation of a second order variable with a term Xt is part of the general framework (and thus not susceptible to re-interpretation) or should be analyzed as an implicit predication relation $p(t,X)$.

See also the discussion in §3.5 of 'innocuous' inscrutability.

discriminate between necessarily equivalent propositions.[33] The focus on coarse-grained propositions is not accidental then: it is a byproduct of the kind of reductive story of mental content that is offered.

The point is that one must be sensitive to the strategic demands of an overall account of intentionality. Appealing to finely discriminated intentional states may well remove the threat of semantic inscrutability, but the problem is relocated to the foundational account of the representational content of mental states.

One might think, for independent reasons (perhaps due to the problems of 'logical omniscience') that belief contents should be represented by structured propositions rather than the course-grained ones favoured by Lewis and Stalnaker. In that case, we can state our problem in one of two ways. (1) We might say that, from a foundational perspective, the kind of 'head-first' stories offered by Lewis and Stalnaker do not deliver genuine belief contents, but only 'proto-belief content', which has less structure. The Lewis-style interpretationist hopes that this proto-content may be enough to found a story of the semantic properties of language; whereupon one might be able to use linguistic behaviour (or the semantic properties of a language of thought) to found true belief content. (2) Alternatively, we might say that the 'head-first' stories do not determine the content of belief precisely; that such an account makes it indeterminate which of a variety of necessarily equivalent belief contents an agent has. This indeterminacy in mental content then bleeds to an indeterminacy in the data to which the interpretationist appeals: for example, it will be indeterminate whether the convention is to utter "Susan runs" only if one believes the content $\langle$Susan, running$\rangle$; or whether it is to utter that sentence only if one believes (the necessarily equivalent) $\langle\phi$(Susan), (running)$^{\phi}\rangle$. Indeterminacy in the data will lead to indeterminacy in best semantic theory once again.

I have argued that the problem of inscrutability is relocated by the move just considered, rather than eliminated; and I have sketched why one kind of supplementary story about mental content (the broadly decision theoretic machinery used by Lewis and Stalnaker) does not give the resources to assign the kind of fine-grained mental content needed to allay our concerns. Nevertheless, relocation may be effective if the problem is more tractable in the new setting. If one had some other head-first account of mental content, or some alternative to the Lewisian convention based pairing of sentences with propositions, then one might be able to resolve the difficulties exactly by providing suitably fine-grained data. Let me sketch one kind of view that, if it could be appropriately filled out, would sustain this case. Suppose that one thought that the data-set for interpretationism should be a pairing of Armstrongian states of affairs[34] with sentences. The pairing would be determined by causal connections between states of affairs and utterances of observation sentences such as 'that's a ball!'. Crucially, if *a's being F* is a state of affairs with *a* and *F* as constituents, metaphysics provides no 'permuted variant' of the state of affairs (states of affairs are not 'abundant' in that way). Such a causal story about the interpretationist's primitive data, together with an appropriately sparse ontology,[35] seems a promising line of attack on the problems we face. Notice the obvious restriction however: it applies in the first instance to medium sized dry goods, and we would owe some other account of the content of other parts of language, such as theoretical and

[33] See Stalnaker (1984), passim. Jeffrey (1965).

[34] cf. Armstrong (1978b).

[35] The details of the metaphysically underpinning will be crucial. For example, a bundle theory of objects (say, objects as bundles of tropes) will not obviously allow the same kind of resolution. Moreover, even on Armstrong's favoured account, macroscopic properties such as 'being a ball' are not universals, and so are not constituents in states of affairs.

In a more general setting, there will be an issue within a causal theory about how to extract objects and properties from the relata of macroscopic causation. If the relata are events, we need to know what privileges the event described as *a's being F* over ϕ*(a's) being (F)*$^{\phi}$. See Kim (1974), Lewis (1986d). This is particularly pressing on reductive accounts of causation, such as Lewis' counterfactual theories Lewis (1973a, 2004), where *prima facie* necessarily co-instantiated events would stand in the same causal relations. One *prima facie* attractive approach on such reductive views is to think that the description of causal relata as involving one object rather than another, should be dealt within a theory of causal explanation, rather than causation proper: if so, an account of such distinctions may well presuppose mental or semantic content rather than explicating it. (This kind of idea is close to that raised and mocked by Fodor (1987, p.126-7) as "alleged interest-relativity of explanation".)

mathematical vocabulary.[36]

Review

A respectable form of attack on inscrutability, consistent with interpretationism, is to relocate the problem to the data, either by postulating structured mental content, or by some more direct means. Distinguish such *relocated* attacks from the *straight* responses, which will not appeal to structured data, but either accept inscrutability or try to deal with it within the story about how coarse-grained data determines theory. The structured-propositions proposal, and its apparent immunity from permutation arguments, may be a contribution to the first project, but it makes no progress on the second.

5.3.2 T-sentences

The framework: Davidsonian truth-theoretic semantics

In the foregoing, we have been concentrating on model-theoretic approaches to semantic theory, which provide valuation functions assigning to expressions in the language a *semantic value*. The Davidsonian setting is somewhat different. The style of semantic theory advocated in Davidson (1967) focuses not on *valuations* and other model theoretic techniques, but on an axiomatic theory from which semantic statements are syntactically derivable. The theorems that associate sentences with truth-conditions are roughly of the form:

> "Londres est jolie" is true iff London is pretty

The data-set that the interpretationist must provide should take the form of a list of such T-sentences. The semantic theory itself then gives axioms concerning the reference of terms and the satisfaction-conditions of predicates.[37] In a popular formulation, the theorems of the theory will be those that are *canonically derivable* from the axioms[38], where canonical derivability is more restricted than logical derivability. In this framework, the T-sentence above will be a theorem of the theory, but the logically equivalent:

> "Londres est jolie" is true iff London is pretty and everything is self-identical

will not be. By the restriction to canonical derivations, so called 'Foster problems' (Foster, 1976) are avoided.

Davidson (1979) is one of the few explicitly to accept permutation arguments for radical inscrutability, in the form given by Wallace (1977). It is somewhat surprising, then, to note that the argument is problematic within the semantic framework that Davidson favours, in just the same way as it is within the Soamesian structured-propositions framework sketched above.

The obvious adaption of the permutation argument to the present case is to adopt permuted axioms of the following form:

> 'Londres' refers to ϕ(London)
>
> 'x est jolie' is satisfied by a iff $\phi^{-1}(a)$ is pretty

On this basis we will be able to give a canonical derivation of the following:

[36] Compare Fodor (1993, ch 4.). A strategy for ascribing structured mental content is proposed by Dretske (1981): close attention should be paid to the assumptions needed to get the fine structure of content determinate.

[37] There will be similar axioms for other categories of expression: see Larson and Segal (1995)

[38] See, for example, Davies (1981); Larson and Segal (1995); Kölbel (2001)

"Londres est jolie' is true iff $\phi^{-1}(\phi(\text{Londres}))$ is pretty

Of course, it is a mathematically trivial move to cancel the permutation and its inverse to obtain the T-sentence originally cited. The restriction to canonical derivations, however, prevents us appealing to just such trivial steps. The T-sentence is not a theorem of the permuted Davidsonian semantic theory, since though mathematically equivalent to such a theorem, it is not itself canonically derivable.[39]

Review

Pace Davidson (1979), the Davidsonian truth-theoretic setting makes permutation arguments highly problematic. As in that case, this formal point should not give one a false sense of security. For, here as before, even if appropriate data sufficiently constrain theory-choice to avoid permutation problems, the problem is pushed back to the selection of the data. Indeed, when Davidson (1980) comes to work out his own views in most detail, settling the content of the language and thought simultaneously, the decision-theoretic framework used builds in exactly the same kind of 'logical omniscience' assumptions that lead to this coarse-grainedness in Lewis' account.[40] *Prima facie*, then, the basis on which the data for interpretationism is constructed is insensitive to exactly the kind of distinctions needed to determine data that would finesse Foster and permutation problems. To avoid the threat, we need further substantive work, whether within or without Davidsonian radical interpretation.[41].

[39] The restriction to canonical derivations is obviously central here, but almost all users of Davidsonian semantic theories wish to avoid "Foster problems" whereby arbitrary tautologies may be added to the clauses governing the satisfaction conditions of predicates. Plausibly, such Foster-problems and the permutation argument stand and fall together, so it is hard to make sense of the position of those (such as Davidson himself) who think that the former but not the latter can be avoided. I do not think that any generality is lost by considering explicitly the canonical-derivability case.

[40] Given the modified Ramsey-Jeffrey setting to which Davidson (1980) appeals, data about ordinal preferences-true fix assignments of cardinal degrees of belief-true and preference-true. Logically equivalent sentences such as "Londres est jolie" and "Londres est jolie $\wedge \forall x(x = x)$" will be assigned the same degree of belief-true and preference-true.

[41] On the latter point, see Lepore and Ludwig (2005, chs. 8, 15.).

5.4 Completeness/compactness arguments for inscrutability

The permutation arguments for radical inscrutability are highly general. They do have one characteristic vulnerability, however, which will be exploited later (§8.4): the deviant models they construct are characteristically *parasitic* on an 'intended interpretation'. I here develop an alternative argument for radical overgeneration, and hence radical inscrutability, which does not have this vulnerability. As we shall see, however, it is harder to argue for the generality that is the principle virtue of permutation arguments.

The theorem and Henkin's proof (details delayed)

Completeness and compactness are well-known metalogical results. They state that if a theory is consistent (/finitely satisfiable) we can build a model for it—an interpretation which renders it true. The Henkin way of proving this result for a first-order theory is to construct an explicit model for the theory. By a variety of sophisticated moves, it extends the theory into one where every existential statement is 'witnessed' by some constant, and which is 'negation complete': containing either S or $\neg S$, for each sentence S. By assigning objects to the constants[42], and including an object within the extension of a predicate only when forced to do so by some atomic sentence within the extended theory, one arrives at an interpretation of the language that (cut down) will serve as a model for the original theory. The Henkin procedure is indifferent to what objects are initially assigned to the constants—hence, by making appropriate choices at the initial stage, we can find models embedding arbitrary reference-schemes.

It is clear how to use this as an argument for overgeneration, and thus radical inscrutability, within the context of a 'global descriptivist' interpretationism that formulates data in terms of a set of first-order sentences.[43] So long as this 'total theory' is not inconsistent, we will be able to build a canonical model by the Henkin methods, embedding arbitrary reference schemes. This is sufficient for radical overgeneration in the context of a standard first order quantificational predicate logic.

To illustrate the general idea, let us consider a toy theory. Consider a constant-free language containing a single non-logical predicate "Dog", and let the theory consist just of the pair of sentences:

$\exists x$ Dog x

$\exists x \neg$Dog x

The theory is compatible with there being infinitely many things—but also compatible with there being just two things in existence. I will show how to find the two element model for this, by means of the Henkin construction.

The first thing to do is to extend the theory T to T', by adding "$\forall x,y,z(x = y \vee y = z \vee x = z)$". Any model of T' will be a finite model of T. Now we run the general Henkin procedure to find a model of T'. Extending the language, we can introduce witnessing constants for the existentials, 'Fido' and 'Betsy', say. We can have as a theory:

$\exists x$ Dog x

Dog (Fido)

$\exists x \neg$Dog x

$\neg$Dog (Betsy)

[42] The assignment is constrained only to assign the same object to any two constants c and c' such that '$c = c'$' features in the extended theory.

[43] Quine in several places, and Putnam (1980) suggest essentially this idea.

$$\forall x, y, z(x = y \lor y = z \lor x = z)$$

It is then clear that we will be able to extend this to a consistent theory T^*, meeting two conditions:

Maximality for every sentence ϕ in the language containing the two new constants, T^* contains either ϕ or $\neg\phi$;

Fully witnessed for each existential $\exists x \phi x$ is in T^*, there is a constant c such that ϕc is in T^*.

Consider the class of all equivalence classes of names, under the equivalence relation "$c = c' \in T^*$".[44] In the case at hand, we can see that 'Fido' and 'Betsy' will be representatives for the two equivalence classes f and b that will result.

The Henkin interpretation of T^* asks us consider the set of equivalence classes of constants as the domain of interpretation—$\{f, b\}$. It then lets each constant refer to its equivalence class—so that "Fido" refers to f and "Betsy" refers to b. Extensions are assigned to predicates in the obvious way: x will be put in the extension of F iff Fc is in T^* for some $c \in x$. In this case, "Dog" will be assigned $\{f\}$. This suffices for a model of T^*, and hence, cut back to the original language, will be a model for the original theory.

To embed another reference scheme, we simply start with some other set in place of $\{f, b\}$, but otherwise parallel the construction.[45]

What is significant, of course, is not that the preceding theory has the model we have just constructed (which we could have found by ad hoc means in a much simpler way!), but that the general technique generalizes to far more complicated theories. The details are involved (particularly in constructing the analogue of T^* meeting conditions (1) and (2)), but the underlying idea is the same. Since this is a standard result, I put the proof in Appendix B.

Higher-order languages

In the context of permutation arguments, we proved the result in a generalized setting. Notoriously, compactness arguments of the kind just considered are sensitive to how the system is set up. Second order logic, with standard semantics, is non-compact, and so we cannot expect to extend the overgeneration argument to that setting.

At first glance, this looks bad for the prospects of arguing for radical inscrutability. For much natural language semantics proceeds on the assumption that higher-order resources are available.[46] One might think, therefore, that this form of argument for radical inscrutability of a natural language is vitiated.

Before moving to a substantive response, two points should be noted. First, there are those who think that semantics can be given in an entirely first-order way. If such a proposal were effective, then overgeneration might be sustained on a set-of-sentences version of interpretationism using just the result cited above. Second, even if overgeneration cannot be proved for the language as a whole, one might still be interested in arguments for radical inscrutability for speakers who restrict themselves to *fragments* of the language. Radical inscrutability looks hardly more plausible for those speaking a more primitive 'first-order' fragment of English.

Still, a less concessive response is available. Though higher order logic with *standard semantics* is not compact, higher order logic with *Henkin* semantics has this property (Henkin, 1950). The difference is this. Standard semantics insists that the domain of second order quantifiers is the full classical powerset of the first order domain. Henkin semantics, by contrast, allows one to restrict this second-order

[44] This will be an equivalence relation thanks to the maximality and consistency of T^*.

[45] Starting with $\{0, 1\}$ gives a 'pythagorean' ontology of numbers of the kind Quine (1964)) considers.

[46] See Partee (1996) for discussion of the liberation effected when Montague introduced higher-order resources into semantic theory.

domain of quantification in various ways. A Henkin interpretation allows one to regard quantificational expressions in a language as restricted in range to, say, the constructible subsets of the first-order range.

Now, it may well be that the Henkin interpretations are 'unintended'. However, from a foundational perspective, the appeal to what is 'intended' cuts no ice: *prima facie* the extent of domains of quantification for any category of expression are up for grabs along with everything else. Hence the liberalism of Henkin models seems quite appropriate to the current setting.[47] The Henkin proofs of compactness for first order logic and a simple type theory are standard pieces of meta-theory: for reference, they are provided in Appendix B.

Matching truth-conditions.

A simple type theory is still *extensional*. Famously, lots of natural language expressions are not: attitude reports, modal vocabulary, and items such as 'seeks', 'is rising', 'allegedly', and 'fake' seem to demand treatment in a richer setting. Two points made above bear repeating: there are some who try to handle such apparently intensional locutions within a purely extensional semantic framework; and there is interest in seeing whether radical inscrutability works within the extensional fragment of a natural language.[48] Clearly it would be nicer if there were a generalization of compactness arguments to an intensional setting. The point is particularly important if the data for interpretationism is formulated in intensional terms—in terms of a pairing of sentences with truth-conditions, rather than with truth-values.

Now, completeness and compactness are concerned with the *truth-values* of sets of sentences, rather than the *truth-conditions* of such. At first it looks like the technique is simply inapplicable to such cases. However, there are ways of *constructing* deviant interpretations that match truth-conditions from the completeness/compactness techniques.[49] Our aim initially will be to associate each expression with an appropriate *Carnapian intension*: a function from indices to compositional extensions appropriate to the category of the expression. We discuss this choice of framework below.

Take a set of sentences that are paired with intensions (functions from indices to truth-values). Now take an arbitrary index *i*. Look at the set of sentences whose intensions map *i* to the true. Call this the *induced theory* at *i* (or, for short, the *i*-set of sentences). Now, we can apply the completeness/compactness results for a type theory to each *i*-set, getting in each case a model which will assign to each expression an extension at *i*. Hence we have described an appropriate Carnapian intension for each expression. By construction, the Carnapian intension assigned to a sentence will be exactly the intension paired with it initially. There are then canonical ways to extract appropriate the compositional intensions that Lewis favours from such Carnapian intensions (leaving the intension of basic categories, and in particular, of sentences, unchanged).[50] We can therefore find a single-indexed general semantics for the set of sentences that matches the pairing of intensions to sentences.

One might have pause at this point. Although every predicate that can be analyzed as having a Carnapian intension might equally be analyzed as having a compositional intension, the reverse is not the case in general. Essentially 'intensional' vocabulary such as 'is rising' and so forth are supposed not to be handleable in the Carnapian way. How have we managed to produce an argument that treats them in this fashion? In fact, the impression that we have given an argument that covers such cases is illusory. Our technique requires us to have a *consistent* theory with respect to each possible situation. However, consider the following triad:

[47] Requiring that the range of the second order quantifiers be 'full' seems analogous to requiring that the first order quantifiers be 'unrestricted'—an assumption usually not built into first-order semantics.

[48] How far this extends is a matter for detailed semantic investigation. See Dowty (1979, ch.4) for an account that treats the progressive 'tense' in English (e.g. 'he is reading a book') as implicitly modal.

[49] Compare Putnam's appendix to *Reason, Truth and History* (1981).

[50] See Lewis (1970a) (pp.196-199 of the version collected in Lewis (1983b)).

- the temperature is rising
- the temperature is ninety
- it is not the case that ninety is rising

These are simply contradictory if ‘is rising’ is interpreted extensionally in a straightforward way. Here we have a limitation of this technique: it cannot be applied to languages containing such predicates.[51]

We have, therefore, a way of building up a ‘Carnapian’ general semantics that matches the truth-conditions assigned to sentences. As before, the generalization to indexical inscrutability is equally secured.[52]

[51] At this point it is important to note that it is controversial whether there are genuine examples of such phenomena in English. (cf. Dowty, 1979, ch.1). Intensional *operators* such as ‘necessity’ ‘always’ and so forth will be fine.

[52] In fact, what we have done is really to replay the simple ‘limiting case’ of the argument for indexical inscrutability described at §5.2, above, where $||^1$ is taken to give an assignment of extensions.

Since we are working with Carnapian intensions throughout, we secure the analogue of indexical inscrutability for indices other than context. By choosing different reference-schemes at different worlds, we can get a great array of possible-world intensions assigned to terms in category *N*. In particular, factors such as the rigidity or non-rigidity of a term will become inscrutable.

5.5 Concluding remarks

As promised, we have developed two distinct kinds of argument for radical overgeneration, in very general settings. We have also developed extensions of those results (most importantly for what follows, to *indexical* inscrutability) and seen semantic frameworks in which the overgeneration arguments do not go through.

What are we to make of these results? In the Gavagai case, we ended up proposing to think of the case as one of a kind of framework inscrutability: as innocuous as the choice of set-theoretic representation of the semantic values of relational predicates. Radical inscrutability of reference is an entirely more shocking proposition. Nevertheless, there have been those that do adopt something like a 'framework inscrutability' attitude to it. Something like this appears to be Davidson's attitude when accepting radical inscrutability. He compares a choice of reference-scheme with a choice of measurement-scale for temperatures:

> To someone who objects [to the inscrutability of reference]. . . the right answer is: individual words don't have meanings. . . . Just as we must indicate whether the numbers we are using to measure temperature place the temperature on the Fahrenheit or the centigrade scale, so we must indicate which method of interpretation we are using.
>
> (Davidson, 1997, p.80?)[53]

Which scale one chooses when giving information is an artifact of presentation, rather than forming part of the information presented. What most would say about the choice e.g. of Wiener vs. Kuratowski set theoretic representations of relations, Davidson appears here to be saying about *reference* in general. Certainly, he regards the inscrutability results as innocuous ones, rather than potential paradoxes to be defused.

Our question in the next few chapters will be: is something like Davidson's attitude sustainable? Is living with radical inscrutability a serious possibility? And if not, what can be done to escape the arguments we have just seen?

[53] References are to the version collected in Davidson (2004).

Part III

Against radical inscrutability

Introduction to Part III

Interpretationism promises a reductive account of semantic properties. Given BEST=FIT, the key notions involved can be spelled out in a clean way. As we have seen, this leads to *radical inscrutability* of reference: there is no fact of the matter about whether "Londres" refers to London or to Sydney.

There is much to gain if we could bite the bullet and accept that reference is inscrutable. Indeed, this is the very attitude that Davidson (1977, 1979) recommends. The purpose of the following two chapters, and this extended introduction, is to examine whether this is a tenable proposal. We shall be looking at the *costs* of radical inscrutability of reference.

It is not my aim here to show where the argument for inscrutability goes wrong, if it does: it is to determine whether or not we are *obliged* to find some flaw in it. To this end, we shall outline five objections to accepting radical inscrutability: the incredulous stare; the alleged self-undermining nature of the arguments; interaction with vagueness; the standing of lexical semantic beliefs; and impact on token inference. We shall briefly discuss these five below. In the two chapters within this part, we develop the final two in more detail: sketching reasons for wanting to attribute beliefs about reference to language users; and arguing for a treatment of token validity that is incompatible with the kind of inscrutability of reference considered earlier.

Five objections to radical inscrutability

The incredulous stare

The thesis under consideration is quite extraordinary. Hold up a red ball in front of your face. Say aloud "that red ball is shiny". Accompany the words by jabbing your fingers into the ball to remove any reasonable way of mistaking your intent. What is being claimed is that nevertheless, the words "that red ball" might just as well refer to the Taj Mahal as to the red, shiny ball in front of you. Isn't that just unbelievable? Moreover, the claims denied seem obvious truisms: 'disquotational' reference principles such as ' "London" refers to London".

For languages in general, and in particular for the reference of one's own language, inscrutability arguments seem to try to deny the non-negotiable. Lewis (1983a, p.46) describes the principle that "Our language does have a fairly determinate interpretation" as "a Moorean fact"—one which we have more reason to believe, than we have justification for believing any premisses that might undermine it.

From time to time, philosophers advocate positions that generate 'incredulous stares'. Lewis himself believed that for each metaphysical possibility, there was an existing concrete cosmos equally as real, and of the same nature as the one we inhabit. Some philosophers ('mereological nihilists') believe that, strictly and literally speaking, there are no tables and chairs, since the simplest particles do not 'compose' any further object.[1] Williamson (1994) argues that there is a fact of the matter about the exact extension of a vague predicate—the number of hairs that it takes to render a man non-bald, for example. In running flat against common sense, radical inscrutability of reference is in the same boat as these doctrines.

[1] See van Inwagen (1990); Dorr (2002); Dorr and Rosen (2002) for arguments for this thesis.

One kind of response to such concerns is to argue that the intuitions can be accounted for within the theory in question. There are many ways in which this tactic could be pursued. I briefly sketch two here: appeal to a substitute notion of 'aboutness'; and appeal to 'disquotational' proposities of truth and reference.

(1) Radical inscrutability of reference need not undermine the notion of a (coarse-grained) proposition's being 'about' an object. Since the propositions assigned to sentences are invariant under the kind of radical inscrutability we have been considering, this may give a derivative sense in which a sentence can be (determinately) about an object, even if there is no fact of the matter at all about what its constituent terms refer to.

According to orthodoxy, John must exist in all worlds where the proposition expressed by 'John runs' is true. As a first approximation, then, we say that a proposition p is ABOUT an object o iff (p holds at w,t $\Rightarrow o$ exists at w,t).[2] (So-defined, the set that is John's singleton, and its singleton, and so forth, also stand this relationship whenever John does. Moreover, as currently defined, any proposition will be ABOUT every necessary existent, contradictions will be about everything whatsoever, and necessary truths about all and only necessary existents. These are undoubtedly problems for the notion just outlined, but should not obscure the fact that for a wide range of cases it gives plausible results: in particular, for contingent characterizations of contingent existents. For the moment, we could either accept these as features of the account, or rule them out on a case by case basis.[3])

Suppose '*Fa*' expresses the proposition p. We should not in general assume that p is ABOUT the referent of '*a*', even on the standard interpretation. 'Beethoven is famous' may well be true relative to the present moment; but Beethoven himself does not exist now. Therefore the proposition expressed is not ABOUT Beethoven (at least in the strong sense). Moreover, let us introduce the monadic predicate 'is a johero'—a person satisfies this iff he or she is one of Jo's heros. The proposition expressed by 'Beethoven is a johero' is then ABOUT Jo, rather than Beethoven.[4]

Since deviant reinterpretations keep propositions expressed by sentences invariant, the proposition expressed by 'John runs' is ABOUT John even if 'John' refers to Jane. The proposition is not ABOUT Jane; since, as discussed in §5.1.2 above, there will be worlds where the proposition is true but Jane does not exist (a world where John runs but Jane is never born). Absent radical inscrutability, if '*Fa*' expresses p, then *often* but not *inevitably* the referent of a and the object that p is ABOUT will be one and the same. Radical inscrutability preserves the ABOUTNESS facts, by renders indeterminate all connections between aboutness and reference.

The inscrutabilist may appeal to ABOUTNESS in attempt to alleviate some of the incredulity that attaches to radical inscrutability claims. They should point out that they are not denying that there is a good sense in which the *proposition* expressed by "London is pretty" is *about* the city London, and not Paris, even if there is no fact of the matter to which city "London" refers. The inscrutabilist might claim that the initial repugnance of radical inscrutability lies in an illegitimate slide from radical inscrutability of the semantic properties of reference, to the (false) conclusion that it will somehow be indeterminate what our beliefs and assertions are about.[5]

[2] We can get alternative notions of 'aboutness' by requiring only that the relevant objects exist at the *worlds* at which the proposition is true, but more 'necessary connections' such as those about to be mentioned would then cause concern—e.g. those arising from putative essentiality of origin.

[3] I hope that the notion could be refined in a principled way to avoid such problems; but there are difficulties, particularly in allowing relational predication 'John is a member of singleton-John' to be ABOUT both John and singleton-John.

[4] Notice that wide-scope tense and modal operators will 'screen off' aboutness. For example, 'it was the case that John ran' is not ABOUT John, since he need not exist for it to be true. But 'John is such that he was running' will be ABOUT John. Similar remarks go for *de dicto* and *de re* modalizing.

[5] This response, however, will be available only to interpretationisms that guarantee invariance of truth-conditions. Global descriptivism, for example, does not take this form.

(2) A different way of addressing the incredulous stare is to claim that one can *agree* with the intuitions, as expressed in ordinary contexts. This kind of response would focus on ordinary *statements* about reference. Within the interpretationist framework, one interprets another's words to make them, as far as possible, turn out correct. Since speakers often assert instances of the disquotational scheme, the onus is on the interpreters to find a way of interpreting the words so they are satisfied. The point is that compatibly with the heavyweight representational notion of REFERENCE suffering from radical inscrutability, there may be a way of understanding all the *object language* vocabulary (in particular, the object-language 'reference') in ways that vindicate ordinary thought and talk. After all, all that is needed is a suitable pairing of objects and words to be assigned as extension to the relevant relational symbol in the target language.

It is important to get in focus exactly what is being proposed. It is being claimed that, as things turn out, words like the French "se réfère" do not pick up on the representational notion that has been our concern throughout this thesis. Rather, ordinary thought and talk latches onto some *other* notion that is better equipped to render true apparent platitudes. We might, for example treat such vocabulary as expressing a relation that is *stipulated* to satisfy the disquotational schema.[6] What the radical inscrutability arguments then show us, we might think, is that such a term cannot express REFERS.

We initiates, then, will be able to diagnose an ambiguity: there is the ordinary lightweight sense of 'refers', in which it sustains the truth of the disquotational principles; and there is the heavyweight sense in which 'refers' expresses REFERS, and which is relevant to the philosophical problem of intentionality.

I think the point is well taken. There is no reason to insist that everyday talk about reference must track the relation that philosophers, for whatever reason, find interesting. As contemporary deflationists about truth urge, the work that can be done with disquotational reference is not to be sneezed at.

We should not overplay the significance of either move for the inscrutabilist. At best, what has been shown is that there is a way to 'explain away' certain intuitions, concerning the notion of 'aboutness' and in favour of disquotational principles. There are situations, however, in which it would be extremely uncharitable to think that people's intuitions against radical inscrutability are directed towards a thin notion: intuitions against radical inscrutability survive transition into the philosophy classroom.

The parallel to other instances of 'incredulous stares' is instructive. Mereological nihilism is not made less counterintuitive if we adjoin to it an explanation of how common-sensical claims such as "there is a table in front of me" come out true. What is found incredible is the philosophical claim that, strictly and literally speaking, they are false.[7] Nor does Lewis' hyper-realism about the existence of concrete worlds avoid its inherent implausibility when its advocate points out that ordinary statements such as "there are no actual talking donkeys" or "there are no existing round squares" come out true on their favoured semantics. In each case, the intuitions against the doctrine are not just semantic intuitions that such-and-such a sentence in ordinary English should come out true. They are focused on the implausibility of the philosophical doctrines themselves.

Nevertheless, the incredulous stare has no *decisive* impact on the question of whether a philosopher should accept radical inscrutability of reference, if she can bring herself to countenance it. It is a 'cost', but if the theoretical benefits delivered are sufficient, it can be outweighed.

[6] Compare McGee and McLaughlin (1994), McGee (2005b) on an 'ambiguity' view whereby the notion of 'truth' falls apart into two notions, both in good order and with theoretical work to do: heavyweight truth and disquotational truth.

[7] Dorr (2002) suggests that the semantic content of such a claim may turn out to be "According to the mereological fiction, there's a table in front of me", or "If there were composite objects, there would be a table in front of me". If so, then semantic content of ordinary existential claims is true, but not ontologically committing. We need a special operator 'strictly and literally speaking' to express our nihilistic philosophical views.

Ineffability

Several philosophers have thought that radical inscrutability renders its own thesis ineffable. In part, this is based on the kind of considerations just raised—that the ordinary notion of 'reference' is used in ways that favour readings on which REFERS is not what is expressed. Even if common-or-garden use of "refers" doesn't pick out the philosophically heavyweight relation, further argument would be needed to show that that relation *cannot* be expressed.

Other reasons may be given, however. If the theorist's language is itself inscrutable in reference, it may be thought mysterious how we conceive of and discriminate between various rival interpretations of our language. In particular, are we not supposing that there is an 'intended' interpretation of the language, from which our 'deviant' interpretations are parasitically constructed? If so, doesn't the conclusion of the inscrutability argument undermine the argument given for it?

I think this kind of dialectical concern is a good one. In fact, for *other* indeterminacy arguments, I think a strong case can be made for the underminingness of the conclusion advocated. Notice that the state in which it would leave us is peculiar. From the *scrutabilist's* point of view, we have no reason to deny that we can discriminate between the rival interpretations. Hence the scrutabilist allows the resources required to make out the argument for inscrutability. *Prima facie*, therefore, we could run the argument to reduce the scrutabilist's position to absurdity. On the other hand, if we maintain that the argument relies on resources that are unintelligible if reference is radically inscrutable, then one who accepts radical inscrutability cannot regard the argument as cogent. We seem to be left with a situation whereby one could be led to adopting an inscrutability thesis, on the basis of an argument that, having accepted its conclusion, one regards as unintelligible.

Such puzzling situations may arise for some kinds of inscrutability arguments,[8] but I do not think that 'undermining' occurs in the case at hand.

The best way of relieving suspicion is to put forward a positive account of how inscrutability theses can be expressed and argued over. In chapter 3 we outlined two ways in which inscrutability of reference for a language can be expressed within an object-language itself—even one innocent of semantic notions such as 'reference'—by exploiting the notion 'Definitely'.

Even if radical inscrutability is not 'ineffable', one might worry that the *argument* cannot be made sense of. Given radical inscrutability, how do we discriminatingly pick out the different 'valuations' in terms of which our discussion proceeded? The appropriate response to this worry differs depending on which style of argument for radical inscrutability is in focus. A compactness-based argument makes no appeal to an 'intended' interpretation of the language, hence is not parasitic in any sense. It does construct a specific 'deviant' interpretation, but unless there is some independent argument that radical inscrutability debars us from performing mathematical constructions, there seems little reason for concern on this point.

Prima facie, permutation arguments for inscrutability do start with an 'intended' interpretation and derivatively specify deviant interpretations. This is, however, entirely inessential. We should simply note that we never need to pick out individual valuations or interpretations—not even an 'intended one'. Rather, we just argue for the *conditional* claim that, for all x, if x is an interpretation-function which renders true a set of sentences T, then the permuted variant of x also renders this true. (Talking of

[8] Consider, for example, Williamson's (2003) considerations for regarding denying unrestricted quantification as self-refuting—since they must allow quantification in a *less* restricted sense in order to specify the restriction. Arguments for inscrutability of quantification based on Skolem-style considerations look to be in danger of undermining themselves in this way. The point is that this seem to presuppose we can in the metalanguage *quantify* over elements not covered by the 'unrestricted' object-language quantifier—which seems only to show that it is not unrestricted.

To emphasize, this shows at most that *accepting* inscrutability results is unsustainable—it does not relieve us of the burden of saying where the argument goes wrong, or even tell us straightforwardly that it is wrong (why should the truth be something that we can have reason to believe?).

different 'interpretation functions' is, of course, just to talk about certain mathematical objects, function from words to objects). This suffices to make the point, for then we need only the principle that there is *some* interpretation which renders the data-set true, to conclude to radical inscrutability. Discussion can thus proceed entirely in *quantificational* terms, without ever mentioning the 'intended' interpretation.[9]

Interaction with vagueness

A more theoretical reason for suspicion of radical inscrutability occurs if we think that the phenomena of vague and indeterminate language are *instances* of the inscrutability of reference—its being inscrutable which precise extension 'red' is associated with, for example.[10]

The problem, alluded to earlier (§6), is that radical inscrutability may undermine various principles used in explaining the phenomena of vagueness. For example, consider the following principles:

> There is someone who is definitely bald
>
> There is no-one who is definitely not bald, such that a man with one hair less would be definitely bald

The truth of these kind of principles (*de re* quantification into definitely contexts) has gained currency as a relatively theory-neutral way of explaining how a non-revisionary logic can be adopted, and still leave room for an account of the seductiveness of sorites reasoning.[11] The idea is that, even though the non-revisionist is committed to the truth of "there is some sharp cut-off point between bald and non-bald men", we can explain the intuitive repugnance of the thesis by supposing that some pragmatic mechanism leads to our hearing such an assertion as "there is some sharp cut-off point between definitely bald and definitely non-bald men".

The problem is that if vagueness and inscrutability are the same phenomena, principles such as

> There is someone who is definitely bald

will be false. For (a) something falls under 'definitely bald' iff it falls under 'bald' on all admissible interpretations; and (b) something which falls under extension of 'bald' on a given admissible interpretation will always fail to fall under the extension of the term on some *other* admissible assignment, given radical inscrutability.[12]

Inference

I contend that *token* inferences of a logical kind are in good-standing *only if the inferrer can legitimately suppose that there is no change in reference during the period in which the inference takes place*. To get an intuitive feel for the kind of problems that might arise, consider inferences involving indexical expressions. *Reiteration* (from S infer S) is about as straightforward a rule of inference as you can find; but it will clearly go wrong in such applications as:

[9] Compare Hale and Wright (1997b).

[10] See Rayo (2004) and Eklund (2005).

[11] See, for example, Fine (1975); Edgington (1997); Keefe (2000); Greenough (2003); Weatherson (2002).

[12] This holds for any predicate P which is not 'universal'—i.e. has at least one object o that does not fall under it in a given admissible interpretation. For then, given an object o' which falls under P on assignment o, consider the permutation that switches o' and o and leaves everything else invariant. By the permutation arguments of Chapter 5, this results in another admissible interpretation, where o' does not fall under P.

Consideration of the 'problem of the many' (Unger, 1980) poses problems for the use of such principles independently of radical inscrutability. In such a case there is available a treatment which will resolve the issue: Lewis' "many" solution (Lewis, 1993). I discuss this case in Williams (2006a).

> The time now is 12 o'clock precisely
> (pause)
> The time now is 12 o'clock precisely

By the time the conclusion is uttered, the time has moved on, and so the reasoning can clearly take us from truth to falsity.

Validity of sentence-type, then, does not secure the good-standing of tokens of that type. In Chapter 7, I will argue that to retrieve a notion of token validity we need to add the presumption that no change in context has occurred that might shift the reference of parts of the sentence. Only with an awareness of type-validity *combined with* knowledge that no relevant change of context has occurred, do we get any kind of logical 'license' for our inference.

We have seen in §5.2 that from BEST=FIT we can argue for an extended *indexical* form of radical inscrutability of reference, with the conclusion that there is no fact of the matter over whether a given expression *retains the same reference* over time. If so, then the supplementary condition on the good-standing of token inference is never determinately met; so no token inference is determinately in good-standing. If the arguments for radical inscrutability are sound, then they generalize to this kind of case. Generalized radical inscrutability ('indexical inscrutability') threatens epistemological disaster, as logical license for inference is debarred.[13]

Semantic beliefs

A traditional view in philosophy is that competence with a sentence can be identified with a cognitive attitude towards the meaning of sentences: perhaps understanding consists in *knowledge of meaning* (Dummett, 1976); or perhaps it consists in *beliefs* about meaning (Heck, 2005a); or some other kind of attitude (Garcia-Carpintero, 2000; Stanley, 2002). Call these *cognitive* approaches to understanding. I will explore whether the cognitive approach would be compromised by radical inscrutability of reference.

Suppose we had an argument to show that understanding a language, under the cognitive conception, required that agents had beliefs *about the lexical meaning* of words. Then we would have the *prima facie* case that the cognitive account is in tension with radical inscrutability. For on this model, to understand words we must believe something of the form:

> *N* refers to *o*

or perhaps:

> *N* refers to the thing which is *F*

Radical inscrutability says that there is never any fact of the matter about whether such claims hold good. *Prima facie*, the upshot will be that one's understanding will consist in beliefs that are indeterminate in status. The *prima facie* case requires the move from 'there's no fact of the matter as to whether *p*' to 'If *x* believes *p*, then there's no fact of the matter whether *x*'s belief is true'. I question this move below. For now, I label it the *infection principle*, and will be assuming it for the time being.

Taking for granted that the infection principle holds, and that the cognitive account identifies linguistic competence with a suitable range of lexical semantic beliefs, what problems arise? Clearly, if the beliefs have to be *knowledgable* then we are in trouble. For under this assumption, by the factivity

[13] Of course, that we are 'logically licensed' to infer something given what we already believe, doesn't mean we *should* so infer. For what we arrive at may be a contradiction—in which case the thing that rationality recommends is to *drop* one of the previous beliefs (or at least lower our credence in them appropriately). If by "belief in Γ licenses the belief ϕ", all we mean is that we should not endorse all of Γ and deny or suspend judgement on ϕ, then this kind of case will not be problematic—it is a matter for separate argument whether on a given occasion we should revise beliefs by modus tollens or modus ponens.

of knowledge, it can be at best indeterminate whether we know the relevant claims, and hence at best indeterminate whether we have what it takes to understand our words.[14] This is surely unacceptable.

However, there are independent reasons for questioning the identification of understanding with *knowledge* of meaning (Pettit, 2002). There is still discomfort with the idea that the attitudes wherein our linguistic competence consists systematically fail to be fully correct, but the primary tension here takes another form.

The basic tension that I diagnose is that the tension lies in the *rationalizing role* for the semantic beliefs that imperil understanding. As I shall outline in Chapter 7, there are grounds for holding that the semantic beliefs that constitute understanding are among one's *reasons* for performing linguistic acts: of uttering *S* when one wishes to say that *p*, for example (Heck, 2005a).

For the third person case, where we can assume that agents are ignorant of radical inscrutability, this role for semantic beliefs does not pose any new questions: *prima facie* false or indeterminate beliefs can rationalize action just as well as true beliefs. The *first person* case is much more disturbing. Any linguistic action, it seems, will commit us to the Moorean paradox:

> *p*, and there's no fact of the matter whether *p*

Believing the first is constitutive of understanding words; but, having come to accept radical inscrutability on philosophical grounds, we are explicitly committed to the latter. One's linguistic actions, it seems, would all have to be undertaken in bad faith.

There may be other roles for semantic beliefs involved in understanding beyond this rationalizing role. They may have an epistemic role, for example, in virtue of playing a part in an inferential model of the epistemology of testimony. Here again, we find potential trouble. I will briefly sketch how this trouble might arise.

Plausibly, the semantic beliefs involved in an epistemology of testimony will not be the *lexical* semantic beliefs whose content radical inscrutability renders indeterminate; but rather the *sentential* semantic beliefs which are in good order. However, on this model of understanding, it is plausible that beliefs about sentential meaning will *derive* from premisses about lexical meaning. (This is, after all, part of the original attraction of looking to semantic theory: to utilize a finite basis of lexical semantic beliefs to explain our competence with potentially infinitely many sentences.)

If sentential semantic beliefs derive from lexical ones, their epistemic standing is imperilled because the beliefs are based on a derivation from premisses that are themselves only indeterminate. Therefore, even though there is no argument against our reaching fully correct beliefs about the truth-conditions of sentences on the basis of our indeterminate lexical beliefs, the status of such beliefs as knowledgable is questionable, since their epistemic heritage is suspect. If the epistemology of testimony appeals to semantic beliefs, and if such beliefs are derived from lexical semantic beliefs that (because of radical inscrutability) are in bad order; then radical inscrutability threatens our ability to gain knowledge through testimony.

I conclude that, if lexical semantic beliefs are involved in a cognitive account of competence, and if the infection principle holds, then radical inscrutability will get us into serious trouble. Under those two assumptions, we will therefore have reason to find flaw with radical inscrutability. Chapter 6 examines the case for these two principles.

Incredulous stares can be faced down; and I have argued that arguments for radical inscrutability are not self-undermining and do not render their conclusion ineffable. Interaction with vagueness is a

[14] Many have assumed that indeterminacy of what is believed is incompatible with the beliefs being knowledgable, but Dorr (2003) puts forward an attractive case for the 'more optimistic' view for indeterminate knowledge where it is indeterminate whether the factivity condition is met.

matter of concern, but it depends crucially on a controversial take on the nature of vague language, and also on the role of ‘Definitely’ in accounting for vagueness, which I will not defend here. The final pair of problems for radical inscrutability will receive extensive discussion, however. Chapter 6 will look at the compatibility of radical inscrutability with accounting for semantic competence or understanding. In Chapter 7, I will discharge the remaining premiss of the argument that indexical inscrutability (and hence by association, radical inscrutability) vitiates the epistemology of inference.

Chapter 6

Lexical semantic beliefs

It is unquestionable that we do have some lexical semantic beliefs: beliefs about what words refer to. It is because we do believe that "Londres" refers to London that we are prompted to stare incredulously at claims that reference is inscrutable. If there is no theoretical role for such beliefs, however, then one may bite the bullet or attempt to 'explain away' such intuitions. Davidson, one of the few to endorse radical inscrutability, explicitly denies that reference has any functional role outside compositional semantics. Reference is:

> a theoretical construct, whose function is exhausted in stating the truth-conditions for sentences
>
> (Davidson, 1977, p.223)[1]

However, if lexical semantic beliefs have serious theoretical work to perform, then holding that such beliefs are universally incorrect will cause trouble.

The claim to be investigated here is that such lexical beliefs are not idle: rather, they are part of what it is to *understand* language. As outlined in the Introduction to Part III, if this is sustained, the costs of accepting radical inscrutability threaten to become overwhelming: our linguistic competence will be constituted by a range of beliefs that are not fully correct; philosophically reflective speakers will be put into Moorean-paradoxical situations; and the epistemology of testimony may be threatened.

I will presuppose a certain view of understanding—the *cognitive account*. This has it that understanding is a matter of having certain semantic beliefs: it is what lies behind the slogan 'to understand something is to know what it means'. We shall not presuppose that the cognitive conception is to be cashed out in terms of *knowledge*, but we will assume that some consciously accessible cognitive attitude such as belief or presupposition is involved. For definiteness, I shall talk in terms of belief.[2]

The plan for this chapter is as follows. The first section elaborates the cognitive conception, dealing with *in-principle* objections, making out an intuitive case for a role for lexical semantic beliefs within the account. I outline a line of resistance: the *weak* cognitive account (suggested by Wright (1987)) whereby only *sentential* semantic beliefs are consciously accessible, and lexical semantic beliefs are merely 'tacit' or subpersonal states.

The second section develops one line of argument for the cognitive conception, giving us a handle on circumstances in which we need to regard semantic beliefs as consciously accessible. This is based on

[1] Page references are to the version collected in Davidson (1984).

[2] See Pettit (2002) for some intriguing Gettier-based arguments against the identification of understanding with knowledge of meaning. I do not find his arguments against identifying understanding with (appropriate) beliefs about meaning so persuasive.

the Davidsonian model of rational language use developed in Heck (2005a). However, Heck's motivating examples do not require more than *sentential* semantic beliefs, so weak cognitivism still seems a stable position.

The third section gives examples where the Heck-style motivation for consciously accessible semantic beliefs transfers applies directly to lexical semantic beliefs. This turns on issues to do with the proper analysis of fragmentary speech acts, which are briefly discussed. The conclusion is that weak cognitivism is unstable; and a proper cognitivism about understanding should attribute consciously accessible lexical semantic beliefs.

As mentioned above, this will be in tension with radical inscrutability only if what we called *the infection principle* holds: if there is no fact of the matter about what *N* refers to, then there is no fact of the matter about whether a belief that *N* refers to *o* is correct. Such a principle is critical in determining the status of any lexical semantic beliefs: if it fails to hold, the case for tensions between radical inscrutability and the cognitive conception of understanding will lapse. As flagged in the Introduction to Part III, the infection principle is not platitudinous: I shall be questioning it in the final section of this chapter.

6.1 The cognitive account of understanding

The cognitive account of understanding, broadly characterized, says that linguistic competence is constituted by cognitive attitudes (say, belief) concerning the semantic properties of language. One way of motivating it, to be explored below, is in the role that beliefs about the meaning of words play in a rationalizing linguistic action.[3]

A more traditional motivation for ascribing 'tacit knowledge' of a semantic theory to speakers is to explain how finite beings are able to attain competence with potentially infinitely many sentences.[4] Since semantic theory displays how one may derive infinitely many 'theorems' about the meanings of sentences on the basis of a finite number of lexical atoms about the meanings of words, it is attractive to say that speakers somehow implement the semantic theory within their cognitive architecture. The natural model is that *beliefs* about sentence-meaning are derived inferentially from beliefs about word-meaning.

Some object that the claim that speakers have a grip on such a complex theory incredible. One move in reaction would be to downgrade the status of the beliefs attributed: to say that the beliefs are 'merely tacit'—subpersonal processing states, rather than consciously accessible person-level beliefs.[5] Such a move, if universalized, would be in tension with the first motivation mentioned above: the idea that semantic beliefs had a role in *rationalizing* behaviour. For beliefs that give a rational explanation of action need to be person-level. All that the 'tacit' semantic beliefs could offer would be *causal* explanations of linguistic action—not what we were after.

There is still general philosophical and psychological interest in the 'downgraded' cognitive conception of understanding. However, for my purposes it is the stronger versions, where at least some semantic beliefs are person-level and consciously accessible, that is of interest.

We can then distinguish two versions of a more ambitious view. The *strong* cognitive conception has it that both lexical semantic beliefs and sentential semantic beliefs are consciously accessible.[6] The *weak* cognitive conception has it that only sentential semantic beliefs are consciously accessible.[7] It is

[3] See, in particular, Rumfitt (1995, §XI).

[4] A *locus classicus* is Davidson (1965).

[5] This is the analogue of a move that Chomsky makes in the context of syntactic beliefs. Dummett (1976) thinks of ascriptions of such beliefs as a 'theoretical representation of a practical ability'—hence mere *implicit*. Evans (1981) introduces the notion of tacit knowledge as an explicitly subpersonal level of representation.

[6] Note this is not to buy into an *ultra-strong* conception, whereby even compositional axioms would be consciously accessible—I take it that really would be phenomenologically implausible.

[7] Note that this is still more ambitious than the *ultra-weak* or 'downgraded' version mentioned above.

the strong cognitive conception that threatens to be in tension with radical inscrutability of reference—for merely subpersonal states would not generate the worries over Moorean paradoxical situations, or the epistemic heritage of testimony, in which the tension could arise. I shall argue in the next section that considerations of the rationality of linguistic acts will favour the strong cognitive conception. For the remainder of this section I shall discuss some initial worries about formulation and plausibility of the cognitive conception.

The regression objection

There is an *in-principle* objection to the cognitive account that I want to dispense with immediately. This is that such an account will inevitably be *regressive*. For in accounting for understanding, we are appealing to semantic beliefs; but what account is to be given as to how we grasp the content of those beliefs—isn't this itself a mode of understanding? At least in this case, understanding cannot be knowledge of meaning, on pain of regress.

The way to resist such an objection is to reject that idea that we need any special explanation of 'grasping the content of a belief' beyond mere *having of that belief.* No correlate of 'understanding' needs to be made out for the semantic belief itself.

Suppose, for example, that our account of belief content is prior to our account of sentence content. Then we must explain directly what it is for someone to have a belief with a certain content—there is no need for a separate account of 'grasping that content'.

On the other hand, suppose that our account of linguistic content is prior to our account of mental content, and assume further that we explicate this along the lines suggested by Fodor and Field. Then thinking is a certain sort of language-use: tokening sentences of 'mentalese' in various cognitive boxes. Believing that 'Londres' refers to London is just to token " 'Londres' refers to London" in mentalese.

Even though this account has us *using* a language of thought, it needs further argument that we need an account of *understanding* for the language of thought. Our use of mentalese might be construed purely mechanistically, with no reflexive awareness at all.[8] To anticipate later discussion: it seems clear that use of natural language is intentional under *verbal* descriptions—there are practical reasoning explanations for why we say what we say using the words that we do. This motivates appeal to beliefs about the meaning or truth-conditions of sentences, that can play a role within the practical reasoning explanations of linguistic action.

By contrast, use of the 'language of thought' does not exhibit the same features. Fodor holds that, faced with a red wall, say, the thought 'that is red' simply happens to us—the mentalese sentence pops into our awareness box (Fodor, 1993). There is little reason to say that our tokening of mentalese 'sentences' is rational under *verbal* descriptions—this is just one way in which a so-called 'language of thought' would be *unlike* a natural language. Whereas we owe an account of natural language understanding—and it looks as though this should proceed in terms of semantic beliefs—there is no corresponding demand for an account of understanding sentences in the 'language' of thought. As in the case of 'head-first' theorists, we simply explain what it is for beliefs to have content directly—there is no residual question over 'grasp' of the belief.

The cognitive conception elaborated

We now put forward an intuitive case for the strong cognitive conception of meaning. On a possible-worlds approach 'London is pretty' and 'London is pretty and arithmetic is incomplete' have the same semantic content. *Intuitively*, though, they are distinct in meaning, so we can expect trouble if we identify

[8] *Conscious* thinking may be another matter; but then we are not committed to analyzing understanding in terms of *conscious* beliefs.

understanding of *S* with knowing that *S* has semantic content *p*, within this framework. Intuitively one can know that a sentence *S* has semantic content *London is pretty* without understanding *S*—for example, if *S* is the conjunction of those words with a complex mathematical truth.[9]

This pushes us to focus on the subsentential level. The strong cognitive conception seems a natural next move: we would say that understanding *S* is to have appropriate beliefs about the content of each expression which forms part of *S*. What distinguishes 'London is pretty' from 'London is pretty and arithmetic is incomplete', after all, is intuitively that the understanding of the former provides only *part* of what one needs to understand the latter. This is not a phenomenon restricted to compound sentences. Consider an invented predicate modifier 'pooky' (of the same syntactic type as 'very'). Contrast the following two sentences:

London is pooky pretty

London is pretty

Now, I can stipulate that 'pooky' is to be a redundant predicate modifier—to satisfy 'pooky *F*' is just to satisfy '*F*'. Given this, the two sentences above will have the same possible-worlds intension. I claim that you don't understand the first sentence just by knowing that it expresses the same possible-worlds intension as the second.

This might be questioned: can we not 'triangulate' from our knowledge of the equivalence between the sentences, to determine the interpretation of 'pooky'? Consider a variant where we introduce a *pair* of modifiers 'smooky' and 'pooky', and tell someone that the sentence 'London is smooky pooky pretty' is equivalent to 'London is pretty'. There are many pairs of assignments to 'smooky' and 'pooky' that might have the effect of 'cancelling each other out', so there's no hope of triangulating to the meaning of either. Nevertheless, we are free to stipulate that, as a matter of fact, they are both redundant modifiers.

The obvious way to deal with this, absent worries about inscrutability, is to say that our understanding of 'London is pretty' is comprised, not only of our knowledge that 'London is pretty' expresses the proposition that *London is pretty* (=the proposition that *London is pretty and arithmetic is incomplete*=the proposition that *London is smooky pooky pretty*); but also of our awareness of syntactical structure of the sentence, and our knowledge of the content of its parts: that "London" refers to *London*, "pretty" applies to pretty things, and so forth. What blocks our understanding "London is pooky pretty" is that we do not know what the meaning of "pooky" is. This seems intuitively satisfying.[10]

[9] See Soames (1989). Contrast Stalnaker (1984) on the 'problem of deduction'.

[10] Soames (1989) argues that the best case for a theory of competence would identify understanding with knowledge of the structured proposition expressed by a sentence. As noted at §5.3, within this framework radical inscrutability of reference would make it inscrutable whether the proposition expressed by 'London is pretty' is:

$$\langle \text{London, being pretty} \rangle$$

or rather:

$$\langle \phi(\text{London}), (\text{being pretty})^{\phi^{-1}} \rangle.$$

An analogue of the infection principle could be formulated for this case, and I think that the dialectic would not much change. (We do not have total inscrutability of the pairing of sentences with propositions, but still enough to undermine the natural thought of identifying understanding with belief that "Londres est jolie" expresses ⟨London, being pretty⟩). Soames' suggestion would therefore simply short-circuit the discussion by allowing us to discern tension between radical inscrutability and the cognitive conception at the level of sentences.

6.2 Rational language use

Given the infection principles, if understanding is to be compatible with radical interpretation, we have to give up on the idea that it is wholly a matter of consciously accessible semantic beliefs. That is, we have to give up on the *strong* cognitive conception of understanding.

We shall now explore whether one can avoid the tension by adopting the *weak* cognitive conception. As flagged in our introductory sections, the issue turns on what we need to do to discharge one motivation for adopting a cognitive conception: to account for what Heck (2005a) calls the 'verbal rationality' of speech. (I shall not be discussing accounts of understanding that do not take this issue on—including the ultra-weak 'downgraded' cognitive conception, as well as non-cognitive accounts of understanding. The arguments below may be regarded as a *prima facie* challenge to those traditions, and so as an argument for a (non-ultra-weak) cognitive conception of understanding.[11])

Let us examine the case for this. Suppose I utter the words "Jill is wearing red". My action is an intentional one, not something that just 'happens to me'. Characteristically, then, we should expect practical reasoning explanations to be available, showing how my beliefs and desires interact so as to make the action 'the thing to do'.

I shall be assuming that the relevant practical reasoning explanations take a familiar Davidsonian form, e.g.:

1. (desire) To make you form the belief that Jill is wearing red
2. (belief) If I say that Jill is wearing red, then you will form the belief that Jill is wearing red.

This belief-desire pair instantiate the classic pattern for rationalizing an action. The assumption is that the beliefs and desires involved have to be consciously accessible if this is to be a good explanation.[12] However, the above practical-reasoning explanation shows only that my utterance is rational under the description *saying that Jill is wearing red* (under what Heck calls 'propositional descriptions'). It does not show that it is rational under the description *uttering the words "Jill is wearing red"*.

Suppose that the utterance *is* rational under verbal descriptions. What kind of practical reasoning explanation should we then give? Plausibly, the following:

1. (desire) To make you form the belief that Jill is wearing red
2. (belief) If I say something that means that Jill is wearing red, then you will form the belief that Jill is wearing red.
3. (belief) "Jill is wearing red" means that Jill is wearing red

With this in place, we can secure the verbal rationality of speech.[13]

We now have a distinction between the propositional rationality and verbal rationality of a linguistic action: the first being the rationality of a speech act under the description (say) *saying that Jill is wearing red*; and the second being the rationality of the same action under the description *uttering the words "Jill is wearing red"*. The second, notice, requires the attribution of semantic beliefs.

The intuitive case for our language-use being verbally rational of language, in addition to its being propositionally rational, is quite strong. Moreover, a more detailed consideration argues in favour of it, drawn from Heck (2005a). Sometimes we utter sentences that betray misunderstandings of the meanings

[11] The most promising approach for such theorists looks to be to challenge the Davidsonian model of rationalizing action presupposed below. This is the line taken by Hornsby (2005) in defence of a non-cognitive conception. Another option would be a Gricean account of non-natural meaning, perhaps as developed in Schiffer (1972).

[12] Davidson (1974) admits no other kinds of belief.

[13] Hornsby (2005) suggests that the relevant practical reasoning explanations involve different ingredients.

of words. Our utterances in such cases can still be *rational.* If I say, for example, "Billy was livid", intending to communicate that Billy was flushed, then there need be no rational criticism of me. Given that "Billy was livid" means *Billy was pale* rather than *Billy was flushed*, there is no obvious way of making sense of the speech act from the point of view of propositional rationality. For the propositional story rationalizes an action of *saying that Billy was flushed*, and no such action took place.[14] Not only does a rejection of the verbal rationality of language seem false to the facts, it also deprives us of the resources to make rational sense of perfectly ordinary speech behaviour. If we make the plausible identification *that which makes language-use rational=that which constitutes understanding*, then we can use this as an argument for the cognitive conception—and indeed, for more than an ultra-weak version of this doctrine, since merely tacit semantic beliefs will not play the required rationalizing role.

We have formulated this as an argument for semantic beliefs in a rich sense—for belief in 'x means that *p*'. Perhaps, so presented, it overplays its hand. Consider, for example, a reformulation in terms of a common belief concerning the truth-conditions, or coarse grained proposition expressed by the utterance :

1. (desire) To make you form the belief that Jill is wearing red
2. (belief) If I utter *u*, and it is common ground between us that *u* expresses the proposition that Jill is wearing red, then you will form the belief that Jill is wearing red.
3. (belief) it is common ground that "Jill is wearing red" expresses the proposition that Jill is wearing red

This seems equally to underpin verbal rationality—but now we need appeal to nothing beyond beliefs about the (coarse-grained) truth-conditions of the utterance. Plausibly, this is the most that the argument as currently formulated can legitimate.

Lexical semantic beliefs

We have seen an argument for the following: beliefs about the truth-conditions of sentences must be of the kind that can subserve practical reasoning explanations. If so, then we cannot think of them as merely 'tacit': they must be the kinds of beliefs that we have conscious access to. However, the argument is *prima facie* particular to the case of sentences. It is enough to secure the rationality of uttering "Billy was livid" when wanting to say that *Billy was pale* that one (mis)believes that "Billy was livid" means *Billy was pale*; there is no obvious need to go further and appeal to a mistaken belief that "livid" means *pale*.

Is there a need for lexical semantic beliefs? Well, certainly it is natural to diagnose mistaken beliefs about the meaning of sentences as *underpinned* by a mistaken belief about the meaning of a word. In the situation above, presumably you would also be inclined to utter "Tony Blair was livid", "my granny was livid" and so forth, in appropriate circumstances where those people were flushed. Within a given agent, sentential semantic beliefs *systematically vary*—which cries out for an *underlying error* that unifies and explains each of these.[15]

Let us grant that the systematicity of errors requires explanation. Still, we have as yet no reason to think that the underlying explanation is a mistaken *belief*, rather than some subpersonal state that is causally responsible for the production of semantic beliefs about sentences—a point forcefully pressed by Wright (1987).[16] In the sentential case, the role of semantic belief within practical reasoning explanations

[14] Heck notes that similar points can be made in the context of ambiguous speech, misheard speech and the like.

[15] See Evans (1981).

[16] See also Wright (1981); Evans (1981); Davies (1981) and Miller (1997).

debarred us from treating it as a merely subconscious or tacit state; and it is *this* that lapses in the current case.

Even if we gave up on consciously accessible lexical meaning beliefs, it is not as though we would be committed to treating sentences as unarticulated 'monoliths'.[17] We can credit speakers with knowledge of *syntactical* structure; and even knowledge that these syntactical parts are relevant to the determination of the meaning of the whole. None of this commits us to lexical semantic beliefs, and none of it is inconsistent with inscrutability.

The Wrightian "weak cognitive conception" thus looks *prima facie* stable. It involves the following claims:

1. We know that *S* expresses the proposition *p*
2. This knowledge is consciously accessible, as is an appreciation of the syntactic structure of *S* and an appreciation that its parts are semantically significant
3. There are no consciously accessible lexical semantic beliefs—rather, there is some subpersonal structure that produces semantic beliefs at the level of sentences as required.

Part of understanding would be the holding of consciously accessible semantic beliefs (i.e. knowledge of the truth-conditions of sentences); partly also a matter of general syntactico-semantical knowledge. But partly it would be a matter of having for each word appropriate neuro-physiological states that, in combination, reliably deliver knowledge of the truth-conditions of sentences. To be sure, we have plenty of conscious beliefs about what words mean; but these are now to be regarded as *post hoc* generalizations, and no part of the ontogenesis of linguistic action.

Verbal rationality and lexical semantic beliefs

The weak cognitive conception of understanding is not without its costs. First, it makes certain connections that one might have thought were *rationally sustained* into merely nomological issues. For example, consider the misunderstanding previously mentioned—manifested in uttering 'Billy was livid' when I wanted to say that *Billy was flushed.* As previously mentioned, sentential semantic beliefs will tend to covary—when I come to realize that 'Billy was livid' means *Billy was pale*, then I will simultaneously realize that 'granny was livid' means *granny was pale*. The view just described represented these patterns as arising from subdoxastic mechanisms that causally influence the sentential beliefs. Now consider someone who continued to hold that 'George Bush was livid' meant *George Bush was flushed*, despite changing all her other sentential beliefs so that 'N was livid' means *N was pale*. By the lights of the present view, this is a mere processing error: there can be no *rational* criticism of the agent. In point of rationality, her beliefs cohere perfectly. This seems odd.

Second, independently of the issues just mentioned, one might think that a right understanding of 'Brutus killed Caesar' should provide the resources to *justify* the right understanding of 'Brutus killed Brutus'. However, on the present view, there is no legitimate *post hoc* reasoning to this effect from the semantic knowledge implicit in understanding.

Third, conscious management of one's lexical beliefs, by looking them up in the dictionary and the like, will become a dubious enterprise. We might think of it as *post hoc* management of one's sentential semantic beliefs via principles that are indeterminate in status—an activity that looks irrational once we become aware of radical inscrutability. Perhaps all that can be said is that by using the dictionary one puts oneself in a situation where one will be caused to have correct sentential semantic beliefs (perhaps in the spirit of Pascal's recommendation that we choose religious friends to improve our chances of acquiring beneficial beliefs.)

[17] *Pace* Heck (cite).

Fourth, second-language competence, if not first, seems inextricably bound up with lexical semantic beliefs: to decide that I should utter "Londres est jolie" in a French oral examination, I explicitly recall the individual meanings of "Londres" and "jolie". Even if I avoid irrationality and epistemic trouble in using English, the worries recur when I speak French or Welsh.

I shall not press these points here, since I think a direct case can be made for lexical semantic beliefs within rational use of our native language. The argument directly parallels Heck's arguments from verbal rationality to the need for consciously accessible sentential semantic beliefs. If successful, they will show that the weak cognitive conception of understanding is unstable—either Heck's considerations should not persuade us to admit consciously accessible semantic beliefs in the first place, or we need such beliefs at the lexical level too.

One often hears it said that 'only a sentence can make a move in the language game'. What is intended, I suppose, is that the typical speech acts have a *sentential* component. Typically, one asserts by uttering a sentence with assertoric force, one commands by uttering a sentence with imperatival force, and so forth. The hackneyed maxim looks false, though. Much of speech consists in 'fragments'. Suppose Billy asks Jean where she will be that summer, and she answers 'In France'. *Prima facie*, she has just performed a speech act whose linguistic vehicle is a *subsentential* expression.

Stainton (1998) gives the following example. We are at a round-table meeting, and I am describing seating arrangements. I do so by pointing at various chairs, while uttering, successively, "the boss", "anyone who needs to go early", "a representative from the faculty board", "Billy". You form appropriate beliefs, respectively that the fourth seat is reserved for the head of department, the second seat is for anyone who needs to go early, and that the first seat is for a representative from the faculty board. Clearly I have managed to communicate propositional contents, but I have done so by uttering fragments of sentences.

If this is right, then we can run the same kind of considerations as before. This sort of language use looks verbally rational. To secure verbal rationality, we need semantic beliefs to play a role in the practical reasoning explanation of the speech act. The semantic beliefs that seem to be required are subsentential: principles such as:

> 'Billy' refers to Billy.

Suppose that I wrongly believe that the chairperson is called 'Billy'. My pointing to the chair and uttering "Billy" is a rational act, with a practical reasoning explanation routing through the (mistaken) belief that:

> 'Billy' refers to the Chairperson.

Again, absent such beliefs, it is hard to see wherein the rationality of the action consists. By the identification of understanding with that which rationalizes linguistic action, we therefore have an argument that lexical semantic beliefs form part of our understanding of language. This is an argument against the weak cognitive conception.

The argument turns crucially on the principle that my utterances are truly *subsentential* speech acts. There is an alternative diagnosis: even though a fragment is all that is phonetically realized, it may be that a sentence is uttered. The parallel is to utterances such as:

> Billy was walking and Jean was too.

Here, the second conjunct is elliptical: there is a phonetically unrealized component. Writing the unrealized component in brackets, we can represent the speech act more fully as:

> Billy was walking and Jean was [walking] too.

In the case at hand, what Stainton (1998) calls 'the ellipsis hypothesis' is that fragments in the kind of speech acts we have been considering are elliptical for full sentences. The suggestion is that phonetic fragments cited above are *syntactically* sentential. For example, the full description of my making the noise "Billy" would be that I utter (something like):

> 'Billy [sits here]'

If the ellipsis hypothesis is sustained, we can give a practical reasoning explanation for phonetically fragmentary utterance that routes through sentential semantic beliefs rather than lexical ones, and the weak cognitive conception would remain an option.

Stainton argues that the ellipsis hypothesis is false,[18] preferring the view that semantic information received from the linguistic fragment is processed *pragmatically* to determine a propositional content for the speech act as a whole. One reason that he gives is that paradigm examples of ellipsis do not occur in discourse-initial position. The following exchange makes clear sense

A: "Mary is at the door";

B: "Billy too"

Here the second sentence is elliptical for "Billy [is at the door] too". Stainton holds that the ellipsis in *B*'s utterance works because the linguistic cue "is at the door" is available in the context in which the phonetic fragment is uttered. Stainton generalizes this to a general account of how a hearer 'fills in' a phonetic fragment to get a full sentence: the hearer must rely on the presence of cues in the prior linguistic context. *Discourse-initial* utterances, of course, will not have any prior cues. Given Stainton's general hypothesis about how ellipsis functions, discourse-initial ellipsis will never be appropriate.

Assuming Stainton's general account of the functioning of ellipsis stands up to scrutiny, it offers a test for putative ellipsis. If the phrase is appropriate in discourse-initial position, it cannot involve ellipsis. Crucially, there are many phonetically fragmentary utterances that can occur in discourse-initial position. (Stainton gives the example of going up to a market trader and saying, *apropos* of nothing, "five red apples, please".) For such examples, Stainton's test rules out the ellipsis hypothesis, and so the argument for a rationalizing role for lexical semantic beliefs goes through.

Stainton takes care to emphasize the provisional nature of the evidence he cites. He maintains, however, that a compelling case can be made for syntactically fragmentary utterances when one considers the evidence as a whole. The issue is a live one within linguistics.[19] The controversy over this point should not bring comfort to the inscrutabilist: for dialectically it is she who will have to adopt controversial positions within linguistics in order to stabilize her position, if she is to resist the arguments via the ellipsis hypothesis.[20]

The weak cognitive conception of understanding is unstable, therefore. We reject the ultra-weak cognitive conception, and accept consciously accessible *sentential* semantic beliefs, because of their rationalizing role. But the same consideration can be replayed to move us to the strong cognitive conception: there is a rationalizing role for consciously accessible lexical semantic beliefs.

The argument with which this chapter is concerned takes the following form:

[18] "it doesn't quack like ellipsis, it doesn't walk like ellipsis, so it isn't ellipsis" (1998, p.854).

[19] See, for example, Merchant (2005) and the references therein.

[20] Moreover, even the fact that such evidence can be brought to bear on the case, and that the evidence just cited appears to tell *against* the ellipsis hypothesis, should be deeply embarrassing for one who wants to stake the tenability of a *philosophical* hypothesis on the ellipsis hypothesis. We have in prospect an *a priori* argument on an empirically substantive point. The tenability of a position such as radical inscrutability of reference should be independent of such empirical matters.

1. Given radical inscrutability of reference, beliefs about lexical meaning are never determinately true.
2. Ordinary linguistic understanding in part consists of lexical semantic beliefs
3. Therefore: Ordinary linguistic understanding consists in part of beliefs in things that are not determinately true.

From this conclusion, we can construct the uncomfortable Moore-paradoxical situations and the epistemological concerns described in the Introduction to Part III.

The focus of this chapter has been on defending (2). Support comes from two directions. First, it is a natural way to work out a cognitive conception of understanding, when one notices that beliefs about 'coarse-grained' truth-conditions of sentences are not plausible sufficient conditions on understanding. Second, principled arguments for adopting a cognitive conception in the first place, on the basis of what is required for speech to be verbally rational, plausibly require not only consciously accessible sentential semantic beliefs, but consciously accessible lexical semantic beliefs. To deny the latter extension, which takes us from a weak to a strong cognitive conception, one would have to adopt tendentious views on the linguistic analysis of fragmentary utterances.

The case for lexical semantic beliefs will only be in tension with radical inscrutability if (1) is sustained. This is supported by the 'infection principle' which tell us that, if there is no fact of the matter about semantic claim p, then the belief that p is at best indeterminate in status. We finish, therefore, with a brief discussion of whether this can plausibly be denied.

6.3 Pseudo-semantic beliefs

The principle at issue holds that radical inscrutability entails that beliefs about lexical reference are never determinately true. The basic idea is simple: if there is no fact of the matter *what n* refers to, then we cannot be totally correct in believing that *n* refers to *o*.[21]

This line of thought, however, presupposes something substantive. The general point, developed in detail below, can be expressed as follows: If the *medium of thought* is itself subject to radical inscrutability, then it may be that the indeterminacy in thought and the indeterminacy in language are so aligned, that we can have determinately true thoughts about reference, despite there being no fact of the matter what either terms of language or the concepts of thought pick out.[22]

Let us *pro tem* suppose that to have beliefs is to token mentalese sentences in one's 'belief box'.[23] A lexical semantic belief about the French "Londres" would be:

> "Londres" refers to London

(Here the underlining highlights the fact that mentalese expressions are here used.) The question of whether the infection principle holds reduces to the question of whether something like the above mentalese sentence can be true.

Just as the natural language 'Londres' is radically inscrutable in reference, it is natural to think that the mentalese term 'London' will be too. There will be no fact of the matter about what either refers to.

Nevertheless, in principle, there may be a fact of the matter about whether or not they *co-refer*. Consider the two names of the famous Roman orator 'Cicero' and 'Tully'. For the inscrutabilist, the reference of such names will each be radically inscrutable. If sentences such as "Cicero=Tully" are to be true, they will need to be *correlatively indeterminate*—co-referring on any optimal semantic theory.[24] Usually we think of such cases as taking place within a single language, but Weatherson (2003) has argued persuasively that we do better to allow correlative indeterminacy across different languages—and in particular, to allow correlations between natural language and mentalese.

Whether or not the mentalese sentence about reference is determinately true depends on how the semantics of the mentalese symbol 'refers' is handled. There is at least one paradigm in the literature—the 'penumbral' treatment of reference of Fine (1975)—which can render it true.[25] One must assume that on a given sharpening, the pairing of a word *w* with *w*'s referent *on that sharpening* will be within the extension of 'refers'. Given this, we have our result: for within the extension of refers will be $\langle$"Londres",$o\rangle$, where *o* is the object (on that sharpening) that is assigned to "Londres". Equally, *o* is assigned to 'London' by the sharpening.

One might worry that the term 'reference' so construed, does not express the robust reference-property in terms of which the inscrutability result is formulated. It is hard to know how to resolve such claims. If we formulate inscrutability in the supervaluational way of §3.1, then the heavyweight notion REFERS simply fails to have application. In this setting, the penumbral handling of 'refers' looks entirely inappropriate. However, if we formulate inscrutability in the purely theory-shadowing way described in §3.2, we get no such quick argument: in the sense there defined there is *no fact of the matter* about the application of REFERENCE. Perhaps there is room for regarding the penumbral handling as appropriate. We need to find a way of reflecting the 'unsettledness' of 'refers' in our candidate semantics.

[21] Rumfitt appeals to beliefs about reference, via their role in rationalizing linguistic action, in objecting to Davidson's view of reference (Rumfitt, 1995, §XI).

[22] The following discussion draws on Weatherson (2003).

[23] See, for example, Fodor (1987, passim).

[24] For discussion of other examples of correlative indeterminacy see Field (1974) and Fine (1975).

[25] See also Field (1998, §I), where in the same way 'correlative indeterminacy' is invoked in giving a semantics for 'refers' and 'is true'.

An appropriate way of achieving this, one might think, is to let the extension assigned to 'refers', on an interpretation, exactly reflect what that interpretation assigns to the first-level terms. After all, *were* that the right valuation, that *would be* the correct extension for 'refers'. Since admissible interpretations disagree on what extension should be assigned to 'refers', it will end up being unsettled what its extension is, exactly as required.

I think that we can avoid taking a position on this dispute. Suppose we had to give up on the claim that the mentalese 'refers' that plays a role in rationalizing linguistic action are genuine beliefs about *semantic* properties of sentences—one might label them instead *pseudo-semantic* beliefs. Whatever status they have, the relevant question is whether they can play the right sort of role in rationalizing linguistic action. And it does seem that beliefs configuring 'refers' can do the job. For example, it is perfectly clear that one could have a *mistaken* belief about lexical meaning, by, for example, tokening the following sentence in one's belief-box:

> "Londres" refers to Paris.

The suggestion on behalf of the inscrutabilist is setting aside the issue of whether (genuinely) semantic beliefs that are part of the ontogenesis of linguistic action; we have pseudo-semantic devices (Fine's penumbral reference) that exploit correlative indeterminacy of thought and language that *match* the inscrutability of natural language. Lexical *pseudo-semantic* beliefs can be granted compatibly with radical inscrutability.

This response to the worry just posed is far from uncontroversial. It requires we take seriously mentalese (already too much for some), take mentalese expressions to have semantic properties of just the same kind as natural language (again, controversial)[26] and then buy into the kind of correlative interlinguistic indeterminacy of Weatherson (2003). Nevertheless, it does give the inscrutabilist a loophole out of the tension argued for above.

[26] Notice that some interpretationisms—for example, Lewis' convention-based approach—cannot naturally be applied to mentalese. There is something ugly about different interpretationist proposals holding good in the mentalese and the natural language case. If parity is enforced, then not all interpretationisms can avail themselves of the suggested defence.

6.4 Conclusion

We started with five arguments against radical inscrutability. One concerned the role of semantic beliefs within an account of linguistic competence. We set out a number of costs that would be incurred if those beliefs which are constitutive of understanding turn out to be merely indeterminate, rather than fully true.

To make the case against radical inscrutability, two claims need to be established: (1) The claim that understanding in part consist of a certain range of lexical semantic beliefs; and (2) The claim that radical inscrutability of reference will imply that such beliefs are untrue.

It is in the second case that we find a loophole: something very like lexical semantic beliefs might be in good-standing *despite* radical inscrutability, so long as the content of thought and of language in the relevant places are indeterminate in correlative ways. Since, on reflection, it is unclear why the considerations in favour of (1) cannot equally be discharged by appeal to these 'pseudo-semantic' beliefs; the case against inscrutability is not decisive.

Chapter 7

Good inference and context

In this chapter, I shall argue that the goodstanding of a *token* inference depends on our being entitled to suppose that reference is stable through the contexts in which we perform the inference.

This 'stability principle' is based on the following line of thought. Treatments of validity for indexical languages typically suffer from a certain sort of incompleteness. The standard Kaplanian treatment of validity in indexical languages requires only that the truth of the premisses secures the truth of the conclusion *within a single context*. This threatens to make mysterious what relevance the validity of an argument has to practical concerns of safely inferring one thing from another: for *token* arguments will typically involve change of context over the course of the inference. In this chapter I outline a modest way of resolving the problem: I hold that particular token inferences are in *goodstanding* when (1) they instantiate a valid argument-type, and (2) we can legitimately presuppose that no change of reference has occurred during the course of the inference. I answer objections to this proposal. In particular, I consider and reject two rival treatments due to Timothy Williamson and John Campbell.

If we endorse arguments for radical inscrutability then we are in no position to resist arguments for indexical inscrutability, which tells us precisely that it is indeterminate whether or not the reference of expressions alter across a given change of context (§5.2). Granted the success of permutation arguments for radical inscrutability of reference, the argument for indexical inscrutability is quite simple: we simply 'string together' different deviant interpretations of the language. Since all the deviant interpretations coincide at the level of the truth-conditions of sentences, a 'cut-and-shunt' interpretation will have the same truth-conditions also. As far as generating the right truth-conditions are concerned, the deviant indexical interpretation is as good as any. If radical inscrutability is acceptable on the basis of such arguments, then, it had better be that indexical inscrutability is likewise acceptable. The treatment of inference argued for in this chapter shows that this is not the case.

7.1 Deductive good-standing

The theme for this chapter is the relationship between what one might call 'speech act' or 'token validity' on the one hand; and 'formal' or 'type validity' on the other. The former concerns the conditions under which the inference from premisses P to conclusion Q is properly made. The latter the study of what follows from what, the implication relationships between the sentences. One might expect the two to be related. Type-validity of the argument from Γ to ϕ guarantees *truth-preservation* in its instances: if the premisses of an argument instantiating the type are true, then the conclusion must be true too. Hence (one would think), in deducing an instance of ϕ from instances of Γ, the validity of the argument-form ensures you won't 'start to go wrong'.

Clearly, however, the study of good inference and the study of what follows from what are distinct. The former is an essentially *epistemological* inquiry. It asks: what are the conditions for a certain kind of epistemically virtuous action? Simple-minded attempts to connect this with formal validity are bound to fail. For example, it will not do to say that, if one believes-true $S_1,\ldots S_n$, and the argument-form from those sentences to S' is valid, then one should believe-true S'. For the appropriate response to noting that S' follows from what was previously believed might be to revisit and revise one's previous beliefs. Further, there are many styles of proper inference that do not seem easily to pair up with logical consequence: inductive inference is a case in point.

To think there is *no* connection between facts about implication and facts about how one should revise one's beliefs would be an implausibly strong claim. It does seem that valid patterns of inference should be epistemologically relevant. The trick is to spell out the relationship.

Here is one view of the relationship that I find attractive. It draws on an analogy with another abstract property/speech act pair: the relationship between truth and proper assertion. Many hold that truth is a (perhaps *the*) norm of assertion. This issues in a connection of the form:

> One must: (assert that p only if p is true).[1]

Similarly, it seems to me that type-validity can be presented as a norm of deductive inference:

> One must: deductively infer ϕ from Γ only if ϕ is a consequence of Γ

Some of the questions one wants to put to this kind of claim have direct parallels in the truth-norm case. For example, note the 'only if' in each formulation: we are not committed to the absurdity that one is obligated to assert every truth of which one is aware, or derive every consequence.[2] Some subsidiary constraints on assertion might be derived from the rule above: perhaps, that one should assert p only if one believes that the conditions demanded by the fundamental norm are met—i.e one believes that p.[3] Likewise, perhaps one should deductively infer ϕ from Γ only if one believes ϕ to follow from Γ. The way in which such 'derivative norms' flow from 'fundamental norms' helps to explain why people may blamelessly assert the false, if under a misapprehension as to what is the case; or blamelessly make a fallacious inference, if under a misapprehension as to what follows from what.

The putative normative relation between inference and implication sketched above is not the only possible model for the relationship between these two notions. The details do not matter, for present

[1] See Williamson (2000, ch.7.) on Assertion for the kind of constitutivef norm here being appealed to.

[2] One can expect many other rules: ranging from constraints of etiquette to the more systematic constraints of Grice (1975). (Note that the above rule is intimately related to Grice's maxim of quality for assertion (do not assert what one believes to be false). There are reasons to assign quality maxim a special role: in particular because, contra Grice on 'conversational implicatures', putative quality-implicatures are not cancellable: attempts at cancellation take the form of a 'Moorean paradox': "p and I don't believe p.")

[3] cf. Wedgewood (2002), Williamson (2000, ch.7).

purposes, for my main purpose here is to consider a worry which, if sustained, would undermine any connection. The attack is a very simple one, arising from standard treatments of (type-)validity for indexical languages. What is interesting is that what seems like a very modest resolution of the problem illustrates the need for a particular kind of stability of content across contexts. It provides grounds for upholding John Campbell's claim (made in a somewhat different context):

> In our reasoning we depend on the stability of language, the fact that its signs do not arbitrarily change in meaning from moment to moment.
>
> (Campbell, 1994, p.82)

As already flagged, indexical inscrutability undermines such a stability principle: so if the best account of inferential practice makes use of it, we have grounds for looking askance at such inscrutability.

7.2 Validity for indexical languages

My concern here is with *semantic* or *model theoretic* validity. At a first approximation, something is semantically valid if *no matter what the subject matter of the premisses and conclusion might be*, the inference is truth preserving. For example, the argument 'Billy is running; therefore something is running' will count as valid in the intended sense, since no matter what "Billy" refers to, or what "is running" means, the truth of the premiss will secure the truth of the conclusion.

This basic idea is cashed out more formally as follows: an argument $p_1,\ldots,p_i,\ldots$, *therefore q* is semantically valid if, on every *admissible interpretation* of the language, either one of the p_i is false or the conclusion is true.[4] Grades of semantic validity will emerge, depending on what we count as an admissible interpretation. If the sole constraint on admissibility is that the interpretation of logical connectives be held constant, as in the "Billy" example above, then we get a notion of formal or logical consequence. More severe constraints on admissibility lead to different consequence relations. The semantic treatment of consequence does not enforce any particular choice of admissibility constraints: what choice is appropriate can be left for independent argument to determine.[5]

So long as we can classify sentences as true or false *simpliciter*, the above definition of semantic validity makes straightforward sense. It is a familiar fact about natural languages that (unambiguous) sentence types can change their truth-value on different occasions of use. We cannot assign a truth-value directly to the sentence "I am George Bush": for it is true when uttered by Bush, and false when uttered by me. Similarly for temporally indexical sentences "it is now 5 o'clock": true as uttered at 5pm, false as uttered an hour later.

Accordingly, even the most obviously 'valid' inference-patterns will on occasion fail to preserve truth, in an indexical language. The rule of reiteration *from S, infer S* is obviously valid for non-indexical languages; but if we instantiate the pattern with "it is now 5 o'clock", and allowing the two tokens of *S* to take place an hour apart, it will clearly fail to preserve truth.

There are two immediate challenges that arise when we consider such phenomena. The first is to characterize a non-trivial notion of semantic validity for argument *types*, for an indexical language. The second is to explain how the notion of validity so-characterized is relevant to the kind of inferences we actually make—token inferences involving potentially distinct contexts.

[4] The notion extends to arbitrary sets of formulae taken as premisses.

[5] Note that in some presentations, the admissibility constraints are obscured, since the semantic theory contains explicit axioms fixing the interpretation of certain expressions. In many presentations of the model theory of first order or modal logic, the semantics of logical connectives is handled by including explicit axioms governing these expressions within the model theory. In a 'general semantic' framework (cf. Lewis, 1970a) logical constants are assigned particular semantic values, and the need for explicit decisions over what should count as an admissible re-interpretation of the language is more evident.

Kaplanian validity

Kaplan's paper "Demonstratives"[6] tackles the first question. The first step towards systematizing validity for indexical languages is to characterize sentences, not in the first instance as plain *true* but rather as *true in the context c*.[7] For example, "it is now 5 o'clock" is *true in the context* c_1 (where the time of c_1 is 5pm) but *false in the context* c_2 (where the time of c_2 is 6pm). We can describe an utterance of the type S as plainly true if the sentence is *true in the context c*, where c corresponds to the setting in which this token of S is uttered.[8]

The basic characterization of semantic validity does not have direct application to languages containing a parameterized truth predicate "true-in-c". At the time at which Kaplan's work was written, there were extant techniques for extending the semantic characterization to parameterized truth-predicates.[9] If we are interested in the *truth-conditions* of sentences—their truth-values in different possible situations—then 'possible world semantics' involves a relativized notion of truth. According to possible world semantics, a sentence such as 'Billy is sitting' is not in the first instance true or false *simpliciter*, it is true or false *at possible world w*. (Again, we can derivatively assign truth *values* to token sentences—as uttered in a concrete situation, S will be plain *true* if it is true *at the actual world*.)[10]

As in the case of indexical languages, the resulting parameterized truth-predicate blocks the straightforward transfer of the characterization of validity in terms of unrelativized notion of *truth*-preservation. In the intensional case, the answer was familiar by the time that Kaplan was writing "Demonstratives": we characterize validity in terms of *truth-at-w* preservation, at all points w, under any admissible interpretation.[11] That is, we first say what it is for an argument to be *true-preserving at all points*: for every point w, either one of the premisses is false-at-w, or the conclusion is true-at-w. As a second step, we generalize over interpretations in the way characteristic of semantic definitions of validity.[12]

Kaplan's "Demonstratives" suggested that we apply the same trick to indexical languages. Again, we find a parameterized truth predicate, truth-at-context-c.[13] To define validity, we must first 'generalize away' this parameter. First, we say what it is for an argument to be *truth preserving at all contexts*: every context c such that the premisses are true-at-c must be such that the conclusion is true-at-c. Then, as a second step, we generalize over interpretations in the way characteristic of semantic definitions of validity. The result is that an argument is valid if it is truth preserving at all contexts, under any admissible interpretation.[14]

[6] Published as Kaplan (1989b), but in circulation in manuscript form since the 1970's.

[7] For present purposes, I ignore other options, such as taking dated *utterances* to be truth-bearers.

[8] One might try to distinguish the *context of utterance* and *context of interpretation*. See below.

[9] The following few paragraphs summarizes some of the setup described in more detail in chapter §3.1 and Appendix A.

[10] As noted previously (Chapter 3), relativization to possible worlds is only one of the ways in which truth-predicates can usefully be parameterized. Others we have mentioned include relativization to variable assignments and delineations (Chapter 3); to times and locations (Chapter 4). See Lewis (1970a, 1980) for general discussion.

[11] It is tempting to gloss this as 'necessary truth preservation' under any admissible re-interpretation. This is misleading, as the usual characterization allows arbitrary *frames* (i.e. re-interpretation of the 'worlds' component of the designated semantics.)

[12] Strictly speaking, we should generalize over models, not just interpretations, where a model is a tuple that specifies, in addition to the interpretation, a 'frame' for the interpretation: sets of points playing the role of context, etc. and any needed accessibility relation on this space. See Chapter 3 and Appendix C.

This way of characterizing validity is 'local' in character, in the sense of Williamson (1994) discussed in Chapter 3 and Appendix C. An alternative to the above would be the 'global' style there characterized, whereby each model for the language contains additional structure corresponding to 'privileged points': which in the designated model correspond to the actual world (for the modal case), the context of utterance (for the indexical case), etc.

[13] I take a context to be a centred world, in the way suggested by Chalmers (1996). See Lewis (1980) for discussion of the utility of centred worlds within a (generalized form of) Kaplanian indexical semantics.

[14] This treatment of validity I regard as the basic idea of the logic of demonstratives. The distinctive theses associated with that work—i.e. that 'I am here now' is a logical truth, and that the necessitation rule fails for logical validity, arise from other features. The first of these is the way that Kaplan restricts admissible interpretations to those that hold fixed the 'character' of expressions (he does this implicitly, by using specific axioms to fix the reference of basic indexical expressions). The second

Kaplan's treatment gives us a definition of semantic validity for indexical languages. Significantly, it shows how such a notion can be non-trivial. For an argument to be valid, it is required only that it be truth-preserving at each particular context. It does not require that it be truth preserving when the premisses and conclusion occur in different contexts. Hence, reiteration can be a valid rule, and we are allowed to dismiss as irrelevant 'counter-examples' to its validity involving change of contexts.

Problems

I will assume that the Kaplanian treatment of indexical languages is the correct way to answer the first challenge: to characterize a non-trivial notion of *valid argument type* for indexical languages. Defining validity in this Kaplanian way leads to an immediate puzzle about the relevance of the validity of an argument type, so-characterized, to the practice of inferring one thing from another.

In the basic case of a non-indexical, extensional language there is no gap between validity of patterns of inference (inference-types) and the goodstanding of concrete exemplifications of that type. Since validity guarantees that the inference in question preserves truth *simpliciter*, the information that inference-pattern I is of a valid type assures us that in any token of type I, the truth of all the premisses ensures the truth of the conclusion.

We find a similar situation for the intensional case. If I is type-valid, then we know that the inference is truth preserving in every possible situation. Obviously, any token of that type with which we would be concerned, would take place *within* a single possible world, so as a special case we have it that such tokens of that type will preserve truth.

In the case of indexical languages matters are not so straightforward. *If* the whole of a token inference occurred in some particular context c, then since a valid inference-type is truth-preserving in every context, it is truth preserving in *the one in question*. The situation would be exactly analogous to the intensional case above. However, token inferences—strings of utterances or inscriptions[15]— *prima facie* involve a series of distinct contexts. It takes time to utter or inscribe successively the premisses and conclusion of a token inference. If the premisses and conclusions occur in distinct contexts, there is no guarantee that the token inference will preserve truth, even if it is of a valid type.

Again, we can illustrate this with the reiteration rule. Consider the inference from "Jim walks" (uttered at c_1) to "Jim walks (uttered at c_2). The inference-type here instantiated is clearly valid, on Kaplan's characterization. At any single context, under any re-interpretation, it will preserve truth. However, we can find an interpretation of the vocabulary that will make the premiss true-at-c_1 and the conclusion false-at-c_2. (Example: Consider the re-interpretation according to which "walks" is an indexical predicate designating *walking* at c_1 and *skydiving* at c_2.) We have *an* interpretation which makes this combination of inference-type and change of context non-truth preserving. Absent further information, one can conclude *nothing* about the truth-preservingness of the actual inference one makes, from the fact that the inference-type tokened is semantically valid in Kaplan's sense.

7.3 The modest proposal

The threat is radical. The validity or otherwise of inference-types threatens to become irrelevant for the practical purposes of reasoning. *Prima facie*, all reasoning involves changes of context—and any such change would seem to render the Kaplanian treatment of validity irrelevant. Our challenge is to explain

special feature is the way that Kaplan combines indexical and intentional logics to define LD-validity. Neither of these special features will be presupposed here.

[15] On a language-of-thought model of reasoning, movements in thought from one set of judgements to another fall within this compass.

the *relevance* of information about validity of inference-types to the rightness or wrongness of specific inferential practice.

One reaction is to conclude that we should be interested in some non-semantic notion of valid inference-type. I think that this would be overhasty. There is a modest solution at hand, that shows how the semantic validity of inference-patterns can still have practical significance.

The modest idea is the following. Call a token inference *good* if (1) it instantiates a valid pattern in Kaplan's sense (2) no relevant change of context occurs during the course of the inference. The contention is that the information that a token inference's good-standing, in the above sense, guarantees that it preserves truth, and hence will justify inferential practice. Since part of the characterization of goodstanding is the validity of the argument type, we explain why information about validity is relevant to correct inferential practice.

Note that the characterization of goodstanding allows *irrelevant* changes of context to take place over the course of an inference. Suppose we make an arithmetical inference. Then, intuitively, no change of context will be relevant to this argument token. Whatever change of context occurs, the argument will be good.

Stability and Constancy

Which token inferences are classified as good will depend on how we cash out the notion of 'relevant change of context'. The basic idea here is that a change of context will only be relevant if it changes the intension of some expression occurring somewhere in the inference. Note that though condition (1) is, by its nature, appreciable independently of what the particular meanings of the words are, appreciation that (2) obtains will essentially depend on what indexical character the words have.

Let $\tilde{C}$ be the set of all contexts involved at some stage of a token inference i. If i is of a valid type, then so long as no relevant change of context occurs, we can be sure that we will not be led from true premises to a false conclusion. For we know, by condition (1), that within a single context this can never happen. Now consider the correct interpretation of the premisses; this will be some $c \in \tilde{C}$. Hence we know that the conclusion is true as uttered in c. However, any legitimate reading of the conclusion will be assessed relative to some context $c' \in \tilde{C}$. By condition (2), we know that c and c' assign the same referents to the expressions that feature in the inference (for otherwise, a relevant change of context would have occurred). By compositionality, an assignment of intensions to the primitive parts of a sentence will fix the intension of the sentence as a whole. Hence we know that the conclusion has the same intension when assessed relative to c and c'. Sentences with the same intension are true in exactly the same situations. Since the conclusion is true in c, it is true in c' also.[16]

The basic idea that *a change of context will only be relevant if it changes the reference of some expression occurring somewhere in the inference* needs more exact formulation. I provide two alternative ways of precisifying this notion, formulated for the general case of an intensional, indexical language.

STABILITY.

> Call a set of contexts $\tilde{C}$ *stable* with respect to a set of expressions E, if the character of the expressions in E assign to each term the same intension at every context in C.[17]

CONSTANCY.

> Call a set of contexts $\tilde{C}$ *constant* with respect to a set of expressions E, if any aspect of context that is used to fix the intension of a term in E is common to every context in $\tilde{C}$.

[16] As presented, the argument assumes that all the premisses are uttered in a single context. Clearly the argument will generalize to situations where the premisses are to be evaluated with respect to distinct contexts within a stable set. Similar considerations will also cover cases where different parts of a single sentence are to be evaluated with respect to distinct contexts.

[17] Here, of course, 'each term' takes wide scope over 'same intension'—i.e. each term keeps its intension stable.

Either stability or constancy could be used as an explication of condition (2); for definiteness, let us require constancy.

Review

The central idea here is to limit the *range of relevant changes of context* that we need to worry about when evaluating a certain inference. If we can legitimately presuppose that context does not change in any of those ways, then the fact that the inference is a token of a valid type will guarantee that truth will be preserved through the inference. An attractive feature is that the only knowledge that we need, in principle, in order to work out what kind of context changes are relevant, will be basic linguistic knowledge about the meaning of the expressions used in the inference.

7.4 Objections and responses

Objection 1.

Constancy is too demanding to play the role envisaged above. For context is constantly changing: utterances of premisses and conclusions are never totally simultaneous, for example. No inference involving any temporal indexical will ever correspond to a constant set of contexts.

Reply 1.

There are several responses available, getting progressively less concessive.

(a) The most concessive response is to grant that no inference involving (for example) temporal indexicals will *ever* be in goodstanding, as characterized above. For recall that the aim was never to extend greatly the range of arguments we could legitimately declare valid. Rather, its goal is to save what we already think we have: to explain how the notion of semantic validity can have practical significance in application to uncontroversially valid patterns of inferences carried out in apparently unobjectionable settings.

For example, correct inference probably plays a central role in the epistemology of mathematics. Mathematical English is, however, just an indexical-free fragment of a broader indexical language ("$2+2=4$ and I am here" is intuitively well formed). As part of an indexical language, mathematical English, though it contains no indexicals, is subject to re-interpretations wherein its terms are treated as indexicals. Token mathematical inferences are therefore under threat. However, constancy, and hence stability, are trivially secured for tenseless, indexical-free fragments of language, such as the language of pure mathematics: this captures the sense in which consideration of indexical re-interpretations of the words seems so *irrelevant* to the good-standing of mathematical inferences.

Beyond this, we are used to the idea that we might need to substantially reformulate arguments in order for them to meet the strictest standards of validity—and reformulation in tenseless language might be one of these cases.

(b) A less concessive response is to preserve the potential goodstanding of token inferences involving temporal indexicals by characterizing the notion of goodstanding in terms of stability rather than constancy. The temporal indexical used in inferring "Today is Tuesday" (at 13.01 hours) from "Today is Tuesday" (at 13.00 hours) does not render the relevant set of contexts unstable, though it arguably renders them inconstant.

(c) Less concessively still, one can question whether the context change involved in the above inference does render the set of contexts inconstant. On one view, contexts comprise a string of factors: the time, place, agent, world, etc. at which the utterance takes place. On this picture, the character of the indexical "Today" would be something like "the day containing the temporal component of context *c*". There is then a relevant difference in context during the above inference; and hence the set of contexts associated with the inference would be inconstant. The current objection tacitly presupposes this kind of setup.

An alternative is to represent contexts as centred worlds, and appeal to auxiliary functions that pick out aspects of those worlds. For example, the interpretation of 'today' would be given by ' $\textsc{day}(c)$'. The 'aspect' of the context that is required to be invariant in the case of the "Today" inference, would be $\textsc{day}(x)$ for $x \in \tilde{C}$. This kind of setup is attractive, if only because it seems hopeless to list once-and-for-all all the possible aspects of context that could be relevant to determining the reference of an indexical expression.[18]

If this is the case, then one can analyze goodness in terms of constancy, while still classing the 'Today' inference as good.

Objection 2.

The account fails to capture the interesting logical or conceptual connections between indexical utterances uttered in different contexts. From 'it is the case that F' we can legitimately infer, at a later point, 'it was the case that F'. This inference is not valid in Kaplan's sense, and so the token will not be in goodstanding, in the current sense.

Reply 2.

It is not my intention to say anything novel about such questions at this stage. As outlined in the response (a), above, the radical threat in view is that the Kaplanian notion of validity turns out to be *irrelevant* to actual inferential practice. Indexicality threatens the application of the concept of validity to *all* token inferences. Faced with this threat, the first job is to secure the goodstanding of the boring, run of the mill token inferences that we all take for granted.

This is not to deny the interest of more ambitious projects that try to explain the distinctive character of inferences involving indexicals that do not obviously instantiate a valid inference type—and indeed, ones whose 'validity', if such it is, *depends* on a change of context. I contend (but will not defend here) that an account of such 'dynamic inferences' can be made out that dovetails with the present account: we can outline a chain of good inferences in the sense of this chapter, that take us from "Here, it is raining" to "There, it is raining" as one drives out of and away from a thunderstorm while keeping track of its location; We can exploit similar stories to account for good and bad cases of basing one's utterance "it was raining" on an earlier utterance of "it is raining".[19]

The account given here is not committed to the success of these more ambitious projects. Nevertheless, the availability of an extension of the approach to these more controversial cases guards against accusations that we are attempting to theorize about good inference using the wrong tools (e.g. materials from the theory of reference, rather than the theory of sense).

Objection 3.

The account fails to cover some obviously valid patterns of inference. For example, consider

[18] Lewis (1980) adopts something like this view, based on, among other things, the consideration just canvassed.

[19] The basic ingredients for this story are the account of inference involving demonstratives outlined in Campbell (1994), and its generalization to a wider class of indexical expressions by Prosser (2005). To convert the 'Fregean' setting to one amenable to Kaplanian treatment, we can appeal to an account of 'dynamic words' based on the treatment of proper names in Kaplan (1990).

the inference from ‘it is now 13.46.24 precisely’ to ‘everything is self identical’. The argument is clearly a good one, in virtue of the consequent being a logical truth, but since the temporal component involved in the antecedent will alter through the course of the inference, it will not be ‘good’ in the sense sketched above.

Reply 3.

To begin with, the now familiar point about the aim of the project being to save the applicability of uncontroversially valid inference types, rather than to give a sense in which more controversial types are valid, is again relevant. Let us set this aside for the moment, however.

In the particular case cited, we might be able to modify the definitions of stability or constancy to get the result the objector wants. We could, for example, demand constancy only of aspects of context that are appealed to more than once in the course of an inference. However, new examples would be constructed making the same point: consider ‘it is now 13.46.24 precisely’ to ‘everything is self identical or it is now 13.46.24 precisely’. Clearly, more and more fine grained notions of constancy could be developed, seeking to capture the idea that there should be no change in aspects of context which are *essential to the validity of the inference*. This looks like a project that will have diminishing returns.

Better to bite some bullets. Let us accept *pro tem* that the inference is not good, in the relevant sense. What bad results follow? Notice that we can characterize a derivative sense in which the inference is well-made. For example, the inference from the null set of premisses to ‘everything is self identical’ is valid. If we know this, and also know that (quite generally) a token argument will never take us from truth to falsity if it is a *weakening* of one that is in goodstanding, then we know enough to explain how we appreciate the truth-preservingness of the original argument. I anticipate telling similar stories, on a case-by-case basis, for other examples adduced.

Objection 4.

You haven’t given what you promised! The notion of semantic validity still has no application to token inferences.

Reply 4.

I have chosen to introduce the neutral notion of the *goodstanding* of a token inference, rather than to describe this status as *validity*. This is because I do not wish to be involved in disputes that seem to me to obscure the central point at issue.

One might think that in cases where a token inference is performed where there is change in an aspect of context that affects the reference of one of the terms involved, it would be a category mistake to apply the notion of ‘validity’ to the inferences in question—evaluation of an argument as valid or invalid occurs only *after* we have ensured against ‘ambiguity’-like phenomena in the expressions used. Alternatively, one might want to characterize an inference from “it is now 5pm” (uttered at 5) to “it is now 5pm” (uttered an hour later) as *invalid*.

I take no stance on the issue, which strikes me as essentially terminological. Depending on which approach one favours, one can construct the relevant notions out of the materials provided–in particular, the neutral notion of a good inference. The key feature is to explain how appreciation of the semantic validity of arguments could be practically relevant; and this has been explained, if it is granted that the good-standing itself is practically relevant.

Objection 5.

We need not think of token inferences as involving change of context at all. The modest proposal is therefore unnecessary: semantic validity will suffice.

Reply 5.

This is a view suggested by Williamson (1997). After describing the Kaplanian treatment of validity for indexical languages, Williamson notes the challenge:

> There is a problem. On this account, the validity of an argument tells us nothing about truth-preservation when the premises and conclusion are instantiated in different contexts. However, inference takes time: we judge the premises and conclusion successively. Thus the account of validity looks irrelevant to the inferences we actually make. (Williamson, 1997, p.652)

Williamson holds that the problem only arises if we focus on the relations between occurrent *judgements*. Plausibly, it takes time to move from judging (or supposing) that premises hold, to judging that the conclusion follows. However, Williamson contends that we should properly be focusing on *belief states*, which can be retained over time. These simultaneous belief states can be considered as instantiating, or failing to instantiate, a valid pattern of inference; and no concern about the involvement of different contexts emerge.

I think two points show Williamson's approach does not get the heart of the issue. Let us suppose *pro tem* that his suggestion handles *cross-temporal* context change. Nevertheless, there are other aspects of context that may still vary between premises and conclusion of an inference. We can bring this out with a simple, though admittedly artificial, indexical expression. Let me introduce the indexical 'blig', which will refer to the primary connective of any sentence or sentence-like structure in which it figures, and to the Taj Mahal otherwise. 'Blig is the Taj Mahal' is true; but 'Blig is the Taj Mahal and Blig is the Taj Mahal' is false.

The point is *the placement of the indexical itself* is potentially an aspect of the context with respect to which the indexical is assigned a referent. Though the example is artificial, it would be most uncomfortable to exclude expressions with this feature *a priori*. Indeed, more standard philosophical examples might well be analyzed as involving expressions of this type: e.g. self-referential sentences or thoughts such as "this very sentence is short", which presumably rely on aspects of context such as *the demonstrated sentence*. A related point is that context can change *within* a sentence, as Lewis (1979c, p.241) emphasizes. If the above proposals are accepted, we look in danger of developing a notion of context and inference that is unable to discharge its other theoretical roles. We need to accommodate such context change, not eliminate it.

Further, the account stands in need of an extension to handle ascriptions of validity to sequences of occurrent judgements, written and spoken inferences, and the like. We *do* evaluate such sequences of events as valid or invalid, and we need a reconstruction of this practice. Since the issue of cross-context inferences has not been taken head-on by the Williamson proposal, we have no ready-made proposal for such cases. Presumably, such ascriptions of validity must be handled derivatively from the primary notion of a valid inference involving belief states—but we have been given little indication how this is to be effected.

Objection 6.

There is an obvious response that the current discussion misses. We should distinguish the *context in which a judgement (or utterance) is made* from the *context with respect to which a judgement (or utterance) should be assessed*. Accordingly, we should think of a token inference as assessed at for a single context—though it is indeterminate which one this should be. Since semantic validity guarantees truth-preservation in all contexts, it guarantees truth-preservation no matter which of these contexts we take to be the actual one.

Reply 6.

This strategy is canvassed by both Campbell (1997) and Williamson (op cit). However, it is easy to find examples where it delivers inappropriate results. Consider an argument of the following form:

> This very second is an even number of seconds after *t*
>
> This very second is an even number of seconds after *t*

Now, if some time has elapsed between the judgement that the premiss holds and the drawing of the conclusion, it is clear that the 'argument' by reiteration should not be taken to be in good standing. However, every choice of context will declare it valid, since a choice of context will determinate a time which will be the referent of 'now' in each case. Assessing the token argument with respect to a single context delivers exactly the wrong result: this is an instantiation of a valid argument type that we want to come out invalid.

The point becomes more dramatic if we use the indexical 'blig', introduced above. Consider the following argument:

> Blig is the Taj Mahal

so:

> Blig is the Taj Mahal and Blig is the Taj Mahal

The premise is unambiguously true (in a sentence with no main connective, 'blig' will designate the Taj Mahal); the conclusion is unambiguously false (in a conjunctive sentence, 'blig' will designate Conjunction). With respect to any *single* choice of context, 'blig' will have to designate a single object; and hence the argument will come out as truth preserving. Again, the argument-type is truth preserving, but no argument token should be described as in goodstanding.

The point is that the separate elements of the token argument—the token premiss and conclusion—call for distinct contexts. We cannot force them into a single context without clear distortion of the intended use of the expressions. There just is no admissible single context for the token inference.

I have drawn attention to an incompleteness in treatments of validity for indexical languages. The standard way to secure a non-trivial classification of indexical inferences as valid or invalid threatens to make a mystery of the relevance of formal validity of inferences to practical concerns of safely inferring one thing from another. The danger has been noticed and addressed before (by Williamson and Campbell) but not satisfactorily.

My proposal is a modest one: it holds that particular token inferences are safe (in *goodstanding*) when we can legitimately presuppose that no change of reference has occurred during the course of the inference.

Nevertheless, if this account is right, something significant has been secured: an explanatory role for stability of *subsentential* meaning—the reference of individual names and predicates—within an account of inference. With this in place, radical indexical inscrutability as formulated in §5.2 conflicts with the best account of good inference.

Part IV

Avoiding radical inscrutability

Introduction to Part IV

In the final chapter of this thesis, I will consider in detail what I take to be the most promising *general* approach to avoiding inscrutability. In this introduction, I shall locate it within a space of possible approaches to avoiding radical inscrutability of reference.

Suppose we adopt the general interpretationist thought that semantic facts are whatever best semantic theory says they are. Given this, in the Introduction to Part III of this thesis I suggested that canonical inscrutability arguments can be seen as having the following form:

SENTENTIAL DATA
The selection of semantic theory is based on data about the semantic values of sentences.

BEST=FIT
The sole criterion of goodness of a semantic theory is fitting this data.

OVERGENERATION
Multiple assignments of subsentential reference assign the same semantic values to sentences

therefore:

INSCRUTABILITY
There is no fact of the matter about which objects terms refer to.

This suggest three ways for the arguments to be resisted, consistently with the overall interpretationist thesis.

Deny that all data is sentential One holds that one can non-circularly identify data constraining the selection of theory *other* than that pertaining to the semantic values of whole sentences.

Deny Best=Fit One holds that the selection of a 'best' semantic theory isn't simply a matter of *fitting* that data (i.e. generating in the appropriate way theorems matching the data, and not generating anything inconsistent with it).

Deny overgeneration One holds that, properly regarded, the data about the sentential meaning is 'rich enough' to avoid overgeneration.

The first and last approaches tackle the problem at the level of the data available to the interpretationist. They do so in different ways: the former concentrates on the *form* of the data; what sort of output of the theory it constrains (one might look to inferential patterns, or classificatory behaviour for raw materials here). The latter concentrates on possible *structure* in that data. The key obligation of such an account would be to make the case that one could identify in a non-circular way appropriately rich data, which could be fed into one of the 'fine-grained' semantic theories mentioned in §5.3 (Davidsonian truth-theoretic semantics, and Soamesian structured propositions). There are several possible directions here—perhaps structure is present in the raw materials themselves;[1] or perhaps there are theoretical constraints operative in extracting data from the distribution of non-semantic facts, which disqualify 'permuted' presentations of the data.[2]

What form and structure the data have will vary depending on the details of the implementation of interpretationism. It is unlikely that we will be able to say anything *general* about the prospects of success. As already noted, the particular implementations canvassed in §1.3—global descriptivism, Lewis' convention-based approach, and Davidson's 'unified theory'—look distinctly unpromising on this front. The attraction of addressing radical inscrutability through denying BEST=FIT is that such modifications impact on *all* forms of interpretationism. If successful, we would avoid radical inscrutability without committing ourselves to anything substantive on the particular implementation of interpretationism at issue. In Chapter 8, I look in detail at one particular way of denying BEST=FIT: David Lewis' work on *eligibility*.

[1] In causal relations between language-users and relatively sparse structured states of affairs, for example

[2] Within a Davidsonian setting, the T-sentence data plays a double role: providing the constraints on semantic theorizing, but also in converting information about sentences held-true to information about belief content. Constraints on belief attribution might therefore impose constraints on suitable data.

Chapter 8

Eligibility

To make the case for radical inscrutability, we built into interpretationism the assumption that the meaning-fixing theory is required only to *fit* the data we provide. We now question this assumption.[1]

Interpretationism holds that the selected semantic theory is the *best* theory of a certain range of data. Quite generally, fit with data is only one among many theoretical virtues, others might include: predictive power, simplicity, explanatoriness, and so forth. It looks as if we have a rich set of resources to play with in selecting the theory which fixes the semantic facts, while remaining faithful to the interpretationist's slogan.

There are theoretical pressures which make the 'null' hypothesis BEST=FIT attractive. If we want a reductive account of semantic properties, we should be wary of appeal to properties of theories that are themselves to be cashed out in intentional terms. Whatever we appeal to we must subject to careful scrutiny to make sure it does not undermine our purpose.

The focus of this chapter is on the theoretical virtue of *simplicity*. Quite independently of his metasemantic views, David Lewis has given an objective analysis of at least one dimension of simplicity (what is sometimes known as theoretical 'elegance'). Furthermore, when we plug this account of simplicity/elegance into an interpretationism, we derive Lewis' response to arguments for radical inscrutability—his notion of a semantic theory's 'eligibility' emerges as a *derivative* theoretical virtue.

Famously, the eligibility response is not a metaphysically 'neutral' account—it requires that we buy into one or another version of what Hirsch (1993) calls 'inegalitarianism' about properties. I shall argue in the second half of this chapter that its metaphysical commitments need to be far stronger. In particular, when combined with the kind of 'ultra-sparse' conception of properties that Lewis favours (a form of microphysicalism), the eligibility response engenders problems worse even than radical inscrutability. Therefore, the eligibility response requires a more liberal conception of property ontology, perhaps involving *emergent* universals.

To begin with, however, I shall discuss a more general question: why be interested in specifically *interpretationist* solutions to radical inscrutability in the first place?

8.1 Side constraints.

In §2.3, I argued that interpretationist's slogan—*the semantic facts are those that follow from best semantic theory*—can be made precise in terms of a *theory shadowing paraphrase*. Not all theory-shadowing paraphrases of semantic discourse are interpretationisms, however.

[1] This chapter is based on material drawn from Williams (2006b)

A fairly common suggestion is that, within a generally intepretationist setup, *causal side-constraints* should be added to a theory of reference.[2] For a semantic theory to be successful, the thought goes, one must not only match data about utterance-conditions of sentences, but one must also assign to basic terms (e.g. demonstratives and proper names) referents to which usage of those terms is appropriately causally sensitive.

I take it that there is no prospect of fitting this account of the selection of the meaning-fixing theory into a broader account of what makes one theory better than another *in general*. In the case of theory choice in science, there seems no way even to formulate an analogous constraint. What is being proposed, to put it bluntly, is an entirely *ad hoc* way of selecting the 'meaning-fixing' theory.[3] Nevertheless, if the required causal relations can be specified in an acceptable way, then we could write down a paraphrase of the general form given in §2.3 that would impose such causal connections. Moreover, such an interaction between causal and interpretationist elements would have its advantages: the causal side-constraints would (in a large range of cases) eliminate radical inscrutability; and the interpretationist setting would address problems of generality that face the causal account[4] as well as more specific technical problems that afflict causal theories of reference.[5]

The mixed account just considered does not count as an interpretationism: we can't reasonably describe it as a view whereby the *best* semantic theory of a certain range of data fixes the semantic facts. This raises the question: Why care about the slogan? I will not offer any knock down objections to this approach, but I am deeply uneasy about it, essentially because it strikes me as objectionably *ad hoc*. In the remainder of this section, I shall put forward some considerations against imposing such side-constraints, which will in turn motivate an interest in interpretationisms as such.

Basic explanations

Given a putative constraint on correct semantic theory, one can ask: in virtue of what is that constraint in force? Consider the constraint of fit with the data, in the case of interpretationism. There one can say: semantic properties are just reflections of best theory of the appropriate data; and fitting relevant data is one way in which a theory can be good. If one appealed to simplicity to rule out deviant interpretations, the same story can be given. The explanation consists in two moves: (1) a general characterization of the nature of semantic facts; (2) the observation that the constraint flows from that characterization.

[2] For an example of one theorist endorsing this view, see Hirsch (1993, pp.108-9).

[3] There may be other ways of getting such causality involved: for example, one might hold that suitable causal relations between terms and objects are a way of providing *data* (pairing singular terms and objects) that constrain theory. This is to deny SENTENTIAL CONSTRAINTS. Another, previously canvassed, is to appeal to causal connections between utterances and structured states of affairs as an attempt to get structured sentential data to feed into interpretationism. This may lead to a denial of OVERGENERATION. Such ideas should be distinguished to the suggestion that causal considerations be built into a replacement for BEST=FIT, which is our topic here.

[4] Causal accounts seem to face problems when departing from the paradigmatic cases of words referring or applying to medium sized dry goods: discourse about the abstract, theoretical terms for the very small and very large; and of the semantic properties of connectives ('and') and other elements of language (e.g. 'very'). See the remarks in the appendix to Field (1972) for discussion of the case of connectives; and the worries in Fodor (1993) about non-paradigmatic cases. Accounting for the semantic significance of *concatenation*, fixed by the compositional axioms of a semantic theory, is another challenge for the causal account (cf. the footnote on p.20, above).

One option, not involving causal 'side-constraints', would be to have an independent causal theory of singular term *reference*, and combine this with an interpretationism about semantic values for other expressions. This might be attractive to one in favour of a fully-fledged "identity theory" for objects for which we have demonstrative contact; but who is worried by the generality problems above. I do not regard this "two-step" proposal as *ad hoc* in the way that the mixed theory is.

[5] The mixed view would resolve so-called '*qua*' problems that face pure causal theories: for even if the required causal connections hold between our utterances of 'horse' and horse surfaces as much as between those utterances and horses themselves, appeal to the constraint to render true "internal organs are parts of horses" plausibly would eliminate one candidate (though we would need to take care to see if Gavagai-style re-interpretations were possible). See Sterelny (1990, ch 6.) for an overview of the *qua* problem.

Now, metaphysical explanations have to come to an end somewhere: call these *basic* explanations. In the case just mentioned, it seems we can legitimately reject a follow-up question: Why is it best semantic theory that fixes the semantic facts? We respond that semantic properties *just are* fixed by best semantic theory.

However, putative explanations that are formulated in gerrymandered terms do not seem acceptable as basic. A general feature of explanation, inside or outside philosophy, is that we should *explain away* such complexity. However, the mixed account of theory selection would leave us appealing to an apparently gerrymandered concept 'best-theory-that-in-addition-meets-causal-constraints'. I diagnose the discomfort with the mixed approach, therefore, as resulting, not from any technical inadequacy, but from the contravention of principles constraining metaphysical explanation. With such a proposal, we cannot evade the question: In virtue of what are *those* the features that give the meaning-fixing theory?

One could resist directly by citing some unifying feature. This is less easy to do here than in general. In the general case, one always has the option of adopting a 'projective' explanation: we say that F's just are the things which meet conditions $C_1, C_2 \ldots$, because that's the way we use to the term 'F', or how we apply the concept of F-ness. This sort of unifying explanation appeals to our classificatory practices: the F's just are the things we call 'F'. The analogous move in the current context would be to appeal to the way that people use the word 'reference', or more generally to the *intentions* of speakers to use semantic terminology in a causally constrained way.

At least if one adopts a word-first strategy, appeal to linguistic *intensions* and *desires* in a unifying explanation will not work. We can semantically ascend, including "N refers to x iff N and x stand in relation C" within the *data* constraining the selection of the meaning-fixing theory, but to appeal to the *content* of referential intentions or stipulations is illegitimate. This seems to me an acceptable version of Putnam's notorious assertion that causal constraints are 'just more theory' (Putnam, 1980)—the causal constraints on reference cannot be more than additional data, if there is no non-circular way of explaining how such causal factors could constrain theory selection.[6] It must be an independent metaphysics of representational properties, rather than a matter of our intentional attitudes, that provides an unifying explanation of why the constraints are in force.[7]

Those in favour of the mixed theory have several lines of resistance to the above. They may reject the principle about metaphysical explanation I appeal to; they may argue that an intentional 'unifying explanation' is acceptable so long as we adopt a *head-first* account to the reduction of intentionality, so that by the time we come to give a theory of linguistic content, we can legitimately appeal to mental content. However, interpretationist forms of theory-shadowing paraphrases allow us avoid worries about gerrymandering entirely. They do this by giving a principled characterization of acceptable constraints on the selection of a meaning-fixing theory: any constraint must be, or be derived from, general theoretical virtues.

[6] I emphasize that a 'pure' theory of reference is not susceptible to this criticism: since it *identifies* the reference relation with such-and-such a causal relation, the theorist can legitimately say "reference *just is* causal relation C", without the appearance of gerrymandering. It is interesting to note that Putnam's (1980) criticism of Field (1972) focuses, not on the notorious 'just more theory' gambit, but on the putative need for an *explanation* of such identities. See also Field (1975) and Putnam (1978b) for further discussion.

[7] Compare Lewis (1984).

8.2 Simplicity and eligibility

David Lewis is committed both to interpretationism (1974a; 1975), and to rejecting radical inscrutability of lexical meaning (1984). Therefore, he must find a flaw in the arguments for radical inscrutability. Lewis' response is to say that *ceteris paribus*, selected semantic theories must be more 'eligible' than their rivals, where for one theory to be more eligible than another is for it to configure more 'natural' vocabulary.[8] The thought is that the 'twisted' reference schemes used to argue for radical inscrutability will typically be very much *less* eligible than the 'standard' ones. The property (running)$^{\phi}$—holding of all ϕ-images of runners—picks out a class of items less naturally unified than the *running* itself. Eligibility constraints thus favour standard valuations over their permuted variants.

What is to favour this kind of move over the imposition of causal constraints? One might argue over the details, for example citing the potential applicability of the eligibility response to cases of abstract reference. However, the deeper worry is that it is a fundamentally *ad hoc* maneuver. Assuming that this is intended to be a version of interpretationism, then we must answer the question *why* eligibility should be a constraint on selecting the meaning-fixing theory.

We can find in other areas of Lewis' philosophy the resources for arguing that the eligibility constraint can be derived from better known theoretical virtues, and so finesse the charge of *ad-hoc*ery. For Lewis, I argue, eligibility is implicit in the theoretical virtue of *simplicity*.[9]

In this half of the chapter I argue that the Lewisian response to inscrutability arguments is composed of three moves. Firstly, correct semantic theory is the *best* one that accounts for the relevant data (interpretationism). Secondly, *fitting* with the data is but one kind of theoretical virtue a theory can have; other virtues, such as *simplicity*, can make one theory better than another (the denial of BEST=FIT). And thirdly, one aspect of simplicity is to be analyzed by appeal to objectively natural ('elite' or 'sparse') properties. We shall see that when we put these together, Lewis' eligibility response follows.

I begin by setting out the second and third components, as they emerge in Lewis' discussion of laws of nature.[10]

Additional junk and scientific laws

Consider the ultimate theory T of microphysics, one which gives accurate predictions of the behaviour of all subatomic particles. Contrast T with the theory T', which is just like T except for the addition of a 'redundant' natural law: one which generates no new predictions about particular matters of fact. (Suppose that it governs the behaviour of particles under nomologically impossible circumstances: what an atom would do if it travelled faster than light, for example. Since no actual particles meet the conditions (we will suppose), both the putative law and its negation are consistent with all the local matters of fact that the world supplies.)

If there were basic facts about the world that did not concern local matters about the distribution of fundamental properties in space and time—for example, if there were robust 'law-making facts' of the kind that Armstrong (1983) postulates—then whether T or T' is the correct theory of the world would be settled by correspondence to reality.

[8] i.e. for its primitive vocabulary to pick out more 'natural' facts.

[9] We could alternatively defend eligibility across the board as an independently motivated theoretical virtue. This would involve separating eligibility from those other elements that I regard as involved in the overall account of simplicity, and reserve the term 'simplicity' only for the latter. If one preferred this route, then one could use the considerations below to argue for the need for the theoretical virtue of simplicity to be supplemented by eligibility, and hence defend eligibility as a non-derivative theoretical virtue. I regard dispute on this point as essentially terminological.

[10] Lewis himself notes the analogy between his treatment of laws of nature and his eligibility response to inscrutability arguments in (Lewis, 1983a), but this is not pursued in much detail.

Some theorists do not wish to postulate law-making exotica, however. For these theorists, the truth-makers for scientific theories must be found in the arrangement of matters of particular fact, rather than in abstruse ontology. Call this view *Humeanism*.

As we have already indicated, T and T' fit the matters of particular fact *exactly as well as one another*. If in order for a scientific theory to be true, given Humeanism, it has just to fit with matters of particular fact, then there is no distinguishing T and T'. Call this the *argument from additional junk*: it threatens to show that Humeanism about laws of nature will make it indeterminate whether or not the redundant law included in T' holds.

The response of the Humeans is that there is more to being a good theory of some range of data, than simply being consistent with that data. Fitting with the appropriate range of data is a virtue of a theory, but there are other considerations besides. Among the additional virtues, for example, are simplicity (how economical and parsimonious the theory is); and strength (how many claims the theory commits itself to, and how much of the data it predicts). For Humeans such as Lewis (1986d, pp.xi-xii), to be the correct scientific theory of a range of data, is to be the *best theory* of that data, where the best theory is one that has the optimal combination of simplicity, strength, and fit.

With the generalized notion of 'best theory' in place, we have a recipe for resolving the puzzle over T and T'. Since T' is just T plus additional junk, it is less simple than T. Moreover, this loss of simplicity is not compensated by any gain in predicative power (strength) or descriptive adequacy (fit). T is the better theory, hence (if these are the only candidates we are to consider) it is the correct theory. The threat of indeterminacy from 'additional junk' vanishes.

Simplicity and Naturalness

What makes for simplicity? As noted earlier, some regard simplicity as fundamentally a projective of our subjective appraisal of a theory. In the current setting, however, this would amount to adding a subjective component to the account of what fixes the laws of nature. Surely this would undermine the objectivity of scientific laws. In response to this worry, Lewis (partially) *analyzes* the notion of theoretical simplicity in terms of objective features of the theory.

The first thing that strikes one about the 'junky' theory T' is simply that it contains an extra axiom. That is, the *theory* is syntactically more complex. To address this, count a theory as simpler if it has fewer, and syntactically less complex, axioms.

This alone cannot resolve our puzzle. Consider, for example, the single property: *being such that T' holds*. Let the predicate 'P' denote this property. Now consider the theory T'', which consists of the single axiom, $\exists xPx$. In syntactic terms, it is clearly simpler than T, and arguably matches it for strength and fit.[11] Obviously we do not want it to be the best overall scientific theory, or the whole enterprise will be trivialized.

We can distinguish between relatively *natural* properties: having spin 1/2; being green; being an animal etc.; from relatively *unnatural* properties: being thought of by somebody, being grue (being green prior to t, blue after t); being the mereological sum of the left half of a human and the right half of a donkey; being such that T' is true. At the limit, we can distinguish the *perfectly natural* or *fundamental* properties from all the rest. Lewis insists that, in evaluating a scientific theory for simplicity, the primitives of a scientific theory must pick out these *fundamental* properties—the basic furniture of the world.[12] Once this is done, we can fairly compare theories according to their syntactic complexity.

[11] See Lewis (1983a). The case is not decisive. For although the single axiom $\exists xPx$ *entails* T', it has little *deductive* power. What we say about it will depend, therefore, on whether we characterize 'power' of a theory through its logical consequences, or through its *entailments*. This should not affect the point about interpretationism being made below, though.

[12] I take it that for Lewis this includes basic metaphysical notions like mereological and logical notions (or constitution relations, if one's metaphysics includes such things) as well as the basic notions of microphysics.

In order for this to contribute to an *objective* (partial) analysis of simplicity in terms of syntactic complexity, we need to postulate an *objective* distinction between elite properties and abundant rubbish (I discuss this below).[13] If the metaphysics stands up, we can cash out at least some of our claims about one theory's being simpler than another in non-subjectivist, non-relativist terms. All else equal, one theory will be simpler than another if the first is syntactically less complex than the second, when each is spelled out in fundamental microphysical terms. In particular, then, we can make the case that T' is less simple than T.

Naturalness and eligibility

Now let us turn back to the case of semantic facts. The situation is analogous to the one facing the Humean about natural laws. Various theories were available, all of which fit the relevant range of basic facts; and an unacceptable indeterminacy threatens. It is attractive to respond just as we did in the case of scientific theories—we hold that *fit* is but one among several theoretical virtues. To be the best account of sentential data, a theory needs also to optimize the other theoretical virtues: in particular, simplicity.

Now recall Lewis' objectivistic (partial) analysis of the simplicity of a theory. Strictly, we look at how syntactically complex the theory is *when spelled out in primitive terms*, where these are constrained to refer to perfectly natural properties. Usually, we would formulate a semantic theory by including axioms such as:

> 'is an atom' applies to something if and only if it is an atom.

To evaluate the simplicity of a theory in Lewis' style, we have to replace the term 'atom' as it is *used* by the theory (i.e. on the right hand side of the above biconditional) by a definition of this property in terms of the fundamental properties of physical science. This must be done, not only for scientific discourse, but also for the general run of natural language expressions: "is red", "is a human", "is running", etc.[14]

Equivalently, though, we might assign a 'degree of eligibility' to each property—a measure that reflects how long a definition of that property would be, if spelt out in primitive terms. The degree of eligibility of a property gives a measure of how much complexity is added to the overall theory by an axiom that uses that property. This means that the syntactic complexity of the original theory, *plus the degrees of eligibility of the properties it uses*, together will sum to the syntactic complexity of the theory when spelled out in primitive terms.

We can then restate the result as follows: a semantic theory will be simpler to the extent that it assigns more eligible extensions to the primitive predicates of the object-language.

What we have reached is exactly Lewis' 'eligibility' response to the inscrutability arguments.[15] First, we have the view of interpretationism as 'best' (i.e. highest-scoring) theory, where *eligibility* is one of the factors relevant to gaining a high score:

> Only if we have an independent, objective distinction among properties, and we impose the presumption in favour of eligible content *a priori* as a constitutive constraint, does the problem of interpretation have any solution at all. ... [C]ontenthood just consists in getting assigned by a high-scoring interpretation, so it's inevitable that contents tend to have what

[13] A classic defense of the need for an objective distinction, and advocacy of one particular form that such a distinction might take, is Armstrong (1978a,b). Lewis (1983a, 1986c, §1.5) argues for the utility of the distinction, and canvasses several forms that it might take, without endorsing any particular account. Armstrong (1989) is a more recent survey and evaluation of the options. van Fraassen (1989a) disputes the entire framework of inegalitarianism concerning properties.

[14] How could this possibly be done? See the following section for discussion.

[15] Lewis states that he owes the idea to Merrill (1980).

> it takes to make for high scores. . . . I've suggested that part of what it takes is naturalness of the properties involved.
>
> (Lewis, 1983a, pp.54-55)[16]

Second, we have the view of eligibility as determined by the syntactic complexity of definitions of a property in perfectly natural terms:

> Physics discovers which things and classes are the most elite of all; but others are elite also, though to a lesser degree. The less elite are so because they are connected to the most elite by chains of definability. Long chains, by the time we reach the moderately elite classes of cats and pencils and puddles; but the chains required to reach the utterly ineligible would be far longer still.
>
> (Lewis, 1984, p.66)[17]

If I am right in this interpretation, then Lewis has the resources to answer the general challenge of *ad hoc*-ery in his appeal to eligibility.

In the remainder of this section I will consider a number of worries. The first three are exegetical, questioning the plausibility of my interpretation of Lewis. I maintain that the derivation of the eligibility response just sketched is of interest independently of any relation to the views of Lewis or anyone else, for it shows how the eligibility response flows from the core interpretationist proposal, rather than being tagged on as an *ad hoc* side constraint. However, I am also prepared to argue the exegetical case, and do so below. After dealing with these issues, I turn to more substantive worries.

Three exegetical worries.

Worry 1: primitive naturalness as a rival account

I claim that Lewis identifies the simplicity of a theory with its syntactical complexity, when spelled out in primitive terms. The passages above appear to support this (e.g. "The less elite are so because they are connected to the most elite by chains of definability"). However, in general Lewis is fairly non-specific about exactly what makes for eligibility. One might think that there are other options that are more plausible.

Here is an instance. Lewis considers taking as a metaphysical primitive the notion of the 'relative naturalness' of one property over another. And indeed, when discussing the eligibility response in his (1983a), he says that it is in connection to language that we 'most need' such a gradable notion of naturalness (p.48)[18]. Perhaps, then, eligibility of a theory is not a matter of its syntactic complexity *when expressed in perfectly natural terms*; it is rather a matter of the relative naturalness of the terms in which a theory is formulated.

I think that this is a dubious view, however. To begin with, eligibility is a relation between *theories*, not among properties directly. What is being contemplated is using the relative naturalness of properties, in effect, to induce an ordering of naturalness among the sets of properties that feature in the theory. This puts substantial pressure on the kind of primitive naturalness required. For example, a mere *ordering* of properties as more or less natural does not look a promising basis for an adequate ordering of *sets* of properties—what we need to get a grip on the relative naturalness of two sets of properties is a *measure* of the naturalness of the properties.

[16] Page references are to the version collected in Lewis (1999).

[17] Page references are to the version collected in Lewis (1999).

[18] Page references are to the version collected in Lewis (1999).

Not all the ways of analyzing natural properties that Lewis (1983a) considers can support such rich materials. Indeed, ontological inegalitarianism, appealing to a theory of sparse universals or tropes, looks able to sustain only the idea of a distinction between perfectly natural properties (those corresponding to universals) and the rest. Though this is sufficient to ground the kind of analysis of eligibility I suggested, it would not ground the highly complex metric seemingly needed for the rival view. Since Lewis is officially neutral between these rival analyses of naturalness, I find the suggestion that his treatment of semantic theories is straightforwardly incompatible with the ontological inegalitarianism just mentioned exegetically implausible.

Indeed, the only view Lewis mentions that does seem compatible with having grades of naturalness—taking the second order relational property *being natural to degree n* as a metaphysical primitive ideology—is independently objectionable within a Lewisian treatment of abundant properties.[19] Lewis treats such properties and relations as set-theoretic constructs.[20] There are however, multiple equally good ways of reducing ordered sets to unordered sets;[21] and so it seems that there will be an irreducible element of arbitrariness as to which sets of sets will represent relations such as *is the father of.* However, if some non-symmetric relation is perfectly natural, to what entity does the primitive second order property 'perfectly natural' or 'natural to degree *d*' attach—the Weiner set theoretic construction or the Kuratowski one? Neither option seems motivated, and no other suggests itself .[22]

Worry 2: new vocabulary

The second worry is that semantic theory is not comparable to case of laws of nature. In that paradigm case, all the terms of theory are constrained to pick out perfectly natural properties, and the theory must be true, so interpreted. However, in the case at hand, there is an element of the theory—the semantic notions—that are not interpreted. Unlike in the paradigm case we are not picking a 'best formulation' amongst a variety of true theories.

The appropriate response here, exegetically, is to point to Lewis' fullest best-system analysis of laws of nature—one accommodating chancy laws (Lewis, 1994a). Lewis thought positing primitive 'chance' facts was objectionable. However, he had no wish to exclude primitive chance talk in formulating laws of nature, in the light of its use within physical science, and in particular within quantum mechanics. His solution was to hold a competition between *partially* interpreted theories featuring an uninterpreted notion of chance. Just as with semantic theory, the constraint could not be that the theories as a whole match the microphysical facts, since there are no chance-facts to match. Instead, Lewis uses an appropriate notion of 'fit' with the facts (together with other theoretical virtues) to pick the successful theory.[23] This treatment of 'new properties' is analogous to the reduction of semantic properties in the case of the designated semantic theory. I contend that it shows that there is no problem of principle in regarding interpretationism as a 'Humean' theory of semantic facts, in the Lewisian sense.

Worry 3: Humeanism

[19] The following objection is due to Armstrong (1986) and Forrest (1986). For extensive and persuasive discussion of the concern, see Sider (1996b).

[20] Lewis (1986c, pt.1)

[21] The famous reductions are by Wiener (1914) and Kuratowski (1921). See §3.5 for a brief discussion.

[22] Taking relations as *sui generis* primitive ontology is one option, if not an attractive one. One would then have to face the concerns outlined in Dorr (2004). On the other hand, if one follows Dorr's recommendations, and avoids non-symmetric relations in ones fundamental ontology and ideology, then naturalness-Primitivism may be an option once more. By this stage, though, ambitions of using it to define 'degrees of naturalness' have been dashed.

[23] cf. Elga (2004) for discussion. I regard Lewis' story as giving a quasi-fictionalist reduction of chance talk to the non-chancy 'Humean mosaic'.

I am suggesting, in effect, that Lewis' attitude to semantic facts is a straightforward extension of his Humean account of laws of nature. Lewis himself, though, did not extend a Humean account to any arena other than fundamental physics (Nolan, 2005, pp.10,86-7). Isn't it implausible that he did so in this case?

The point that Lewis is concerned primarily with a Humean account of fundamental laws of nature is well taken: the question is whether there is any obstacle in principle to extending this to the special sciences. Indeed, the factor just mentioned, the introduction of novel vocabulary—marks a major difference between Humean accounts of laws as they were formulated in Lewis (1986d), around the time when Lewis was setting out his metasemantic views. As just discussed, given the later development of Lewis (1994a), this no longer marks a departure.

More substantively I do not see why his account should be in tension with the broader project of Humean accounts of the laws of special science more generally, given by the best theory (possibly involving new properties) of a range of independently legitimated data. In particular, an account of the laws of special science could be fed into Lewis' theory of counterfactuals (Lewis, 1973b) to deal with 'counter-legal' conditionals that are otherwise highly problematic.[24]

Indeed, Humeanism might be seen as more attractive in the context of the laws of special sciences, than in the context of the fundamental laws of nature. Armstrong (1983) favours robust law-making facts in microphysics, but equally he adopts a sparse conception of facts, and is not friendly to the idea of emergent facts special to psychology, economics or semantics that would be needed to generalize the law-makers account of laws to special sciences. Equally the Armstrongian objections to Humeanism seem less pressing in these contexts.

[24] The problem is that Lewis cashes out counterfactuals in terms of similarity between situations, and similarity partially in terms of laws. If the only laws around, then, are laws of microphysics, it will be difficult to account for such true counterlegals such as 'If cows were as small as atoms, we wouldn't be able to see them'. However, if we have an account of the laws of biology, we can stick with Lewis' original law-based analysis of similarity, and handle such phenomena. (There are, of course, more recalcitrant counterlegals, such as "were gravity to obey an inverse-fourth law...". A partial account of counterlegals is far better than no account at all).

8.3 Six objections to Lewis' eligibility response.

In this section, I shall describe and briefly respond to six objections to the eligibility response to inscrutability. Not all of these have satisfactory answers. We shall, in the next section, develop in more detail a worry which gives a more precise focus to the most serious of these concerns, and suggest a alternation of Lewis' framework that will allow the difficulty to be resolved—at a cost.

Objection 1: Inegalitarianism Lewis' response presupposes an ontology of 'perfectly natural properties'. This 'inegalitarianism' is unacceptable. (van Fraassen, 1989a, 1997).

Reply 1: To undermine Lewis' eligibility response, one would have to argue either that there is no division between more or less natural properties (what Hirsch (1993) calls property 'egalitarianism'), or to say that such a division is mind- or language-dependent, so that invoking it would vitiate the reductive metasemantic ambitions. (One could also object to the idea of *perfectly* natural terms—see the next objection).

Beyond these minimal conditions, the eligibility response can be neutral on the kind of property inegalitarianism being invoked. It might be unpacked ontologically—in terms of sparse universals or tropes; or in terms of additional ideology—primitive naturalness of properties, or primitive contrastive resemblance between individuals.

Lewis himself cites all the above treatments. However, if our interest is solely on the eligibility response, we can be neutral in several areas where Lewis is committal. We can hold that the natural properties have their nomic or causal roles essentially for example, instead of holding Lewis' view whereby it is possible for any property to play any role Lewis (forthcoming). We can hold that sparse property ontology consists of world-bound entities, rather than trans-world individuals.[25]

Objection 2: The onion objection Even given a distinction between more or less natural properties, are we entitled to the assumption that there are *perfectly natural* properties?

Reply 2: Consider the following scenario, derived from Armstrong (1978b, ch.15). It is familiar that atomic nuclei are composed of protons and neutrons, and these in their turn are composed of quarks. What if this process of decomposition went on for ever? What if every property P were decomposable into more basic properties Q, Q' in the way that *being a hydrogen atom* is decomposable into properties such as *being a proton* and *being a neutron*? Call such a scenario an *onion world*.

Would there be *perfectly natural* properties in an onion world? One line of thought would have it that there each layer of microstructure is *more natural* than the one above it. If this is the case, then there would indeed by no 'most natural' properties.

This description of the case is not mandatory, however. Suppose we think of the case as one where every universal is structural, composed in the relevant way from universals at the level below. Each universal is then 'infinitely structural' (compare Lewis, 1986a). Alternatively we could have each 'level' featuring emergent universals—simple universals that perhaps nomologically covary with those at 'lower levels'. It seems quite motivated, when we come to read off property naturalness from the ontology of universals, that we count all properties that are 'united' by a universal

[25] For discussion of Lewis' views on properties, and alternatives, see Lewis (1986c), Lewis (typescript) and Schaffer (2005)

If our concern is with linguistic content alone, in the context of a head-first theory, then we could even appeal to attitudinal states in characterizing natural properties without immediately introducing circularity. However, Lewis (1983a, 1994b) also wishes to appeal to eligibility within an account of mental content, which would end up blatantly circular. (For an alternative take on mental content, in terms sympathetic to Lewis, see Stalnaker (1984).)

(whether simple, complex or structural) as perfectly natural in the relevant sense. If so, onion worlds have a relative abundance of perfectly natural properties, rather than none at all.

Objection 3: Can we give an adequate explication of the notion of syntactic complexity?

Reply 3: The natural first move would be to count the number of connectives involved in the theory, axiomatized as a single sentence (this will have application, therefore, only to finitary theories).[26] Sider (1995) objects (1) that disjunctions should count for more complexity than conjunctions (2) that there may be no way of singling out the privileged class of connectives allowed in the definitions. The latter point does not seem over worrying. Two-place connectives and the array of logical quantifiers (perhaps of higher orders) seem perfectly adequate to the task at hand.

The former point seems to presuppose that the assessment of a theory as more or less eligible should track the intuitive idea of a class of properties as overall more or less natural. I think, however, that we should divorce eligibility from naturalness. The eligibility of a property is a measure, not of its naturalness, but exactly of the complexity it adds to a theory. The case for counting disjunctions as detracting from eligibility more than conjunctions relies on the conflation of these two issues.[27]

Objection 4: Optimality A general worry about Humean accounts is that they appeal, not only to individual virtues like simplicity, strength and fit, but also to a notion of an 'optimal combination' of those to arrive at the overall 'goodness' of a theory. What is this additional 'optimizing' primitive and how is it justified?

Reply 4: The general worry affects all Humean accounts of laws.[28] One concessive move here is to relativize the sense of 'best' involved. We would be unable to say that a given theory was best simpliciter, but only that it was best relative to a particular way *w* of balancing virtues. Likewise, we would have a plentitude of possible paraphrases, playing on the various relativized notions of 'best': and the utility of the paraphrases is then a matter of which discharges the appropriate theoretical function. One would then hope that much the same facts would be vindicated on all but crazy ways of balancing the virtues, so that particular choices wouldn't matter much.

The 'realist' alternative is to treat 'best theory' as a basic kind of 'structured' virtue, with simplicity, power, fit etc. as its structural parts. After all, there are many cases where a 'compound' virtue intuitively arises from the balancing of component virtues (e.g. on the weighting of considerations on an objective list account of morality) and it seems a major step to declare every such case infected with subjectivity. This realism allows us to formulate Humeanism, explain the relevance of eligibility and the other theoretical virtues to the determination of best theory, and avoid accusations of subjectivity or circularity. The realism comes in reductive or non-reductive versions. A non-reductive version would hold that there is no informative description of the way in which

[26] Notice also that we should not allow arbitrary *n*-adic connectives in formulating the theory, else we could reduce complexity simply by using single rich connectives. The solution within the spirit of the eligibility response is to postulate 'sparse' universals for logical connectives as well as for non-logical properties, and demand the theory be formulated entirely in ontologically privileged terms. Thanks to Kit Fine for pressing me on this point.

[27] A conflation, it has to be said, that Lewis does much to encourage. Nevertheless, I maintain that the best reading of Lewis' work separates these two notions—see the exegetical comments above.

[28] The introduction to Lewis (1986d) considers this worry and declares it an unattractive feature of Humeanism about laws. One idea there seems to be to treat optimality as fixed by our dispositions to treat something as optimal. If so, then mental content (specifically, regarding combinations of virtues as optimal) are presupposed in any broadly Humean account of laws. This may be a worry for interpretationism, though once again a head-first theory may help out. It is a hugely more difficult challenge within the overall Lewisian programme of 'Humean Supervenience' (Lewis, 1986d, introduction), since mental content is not supposed to be involved in picking out laws of nature (and so the counterfactuals and related notions that are supposedly used to fix mental content).

the structural parts are balanced to determine the overall virtue. The overall virtue is therefore *sui generis*: a strictly additional metaphysical commitment. A reductive version would attempt an informative characterization, perhaps drawing on related work within the philosophy of science in relating simplicity and fit in the evaluation of curve-fitting.

Objection 5: Vagueness Vagueness pervades natural language. Absent ontic vagueness (in disrepute through much of the philosophical community) how can a basis of precise perfectly natural properties allow one to define the extension of a vague predicate?[29]

Reply 5: There are two worries that can be extracted from the above question. The first questions whether we have any hope of combining the eligibility response with a treatment of vague language. The second grants that there are treatments of vague language consistent with the eligibility response, but holds that the eligibility response pushes us towards unattractive treatments.

The first objection presupposes that the perfectly natural properties are precise. If the perfectly natural properties are, as Lewis holds, found at the level of microphysics, this is plausible.[30] However, if the distribution of universals is more liberal—if there are universals at a macro-level, for example—then the situation is less clear. I shall not press this point, however, as it is somewhat unclear what a semantics for a language involving reference to vague properties would look like.[31]

On an epistemicist approach to vagueness, such as that advocated Williamson (1994), there is similarly no problem of mismatch between the range of basic properties and a vague language. For the epistemicist, a vague language can have an absolutely precise semantics. Another option is to analyze vagueness as a kind of inscrutability. In that case, all the semantic theories will be perfectly precise, and the vagueness will consist in indecision over which is selected. Some cases of 'vague language', broadly construed, are easily handled within this setting (e.g. the "mass" case of Field (1973)) but more paradigmatically vague terms (degree vague examples such as 'red', 'bald' and 'heap') are more difficult. For progress towards a metasemantic treatment of such terms, see Rayo (2004).[32] The final response would be to develop a special semantic treatment of vague language, perhaps in the sophisticated supervaluational setting of Fine (1975) or Kamp (1975). In the present context, these would need to be formulated ultimately in perfectly precise terms.[33]

There are, therefore, extant approaches to vague language compatible with the framework of the eligibility response. The most that the present setup could be accused of is committing itself to one or another of a variety of ambitious positions in the philosophy of vagueness.

The second line of objection is interesting. A sophisticated supervaluational semantics for a vague language is going to involve much greater complexity than the straightforward classical semantic theories that correspond to its 'sharpenings'. Given this, it looks as if on grounds of eligibility, classical semantic theories will always do better than their more sophisticated rivals.[34]

[29] Thanks here to Elizabeth Barnes for pressing this point.

[30] Though see Barnes (2005).

[31] Two thoughts: first, one might appeal to (metaphysically) vague sets. Or, one could formulate semantics in terms of precise sets, and pick out certain extensions as 'basic' if they 'precisify' the metaphysically vague property, and then run the eligibility story with 'basic' in place of 'perfectly natural'.

[32] Lewis (1969) takes a 'metasemantic' approach (see the discussion in the appendix to Lewis (1970a)). Burns (1991) tries to defend this approach, and is criticized in Keefe (2000, ch.6.). Eklund (2005) compares semantic and metasemantic treatments of vagueness, and Rayo (2004) develops a metasemantic proposal explicitly aimed at underpinning vague language, using Lewisian notion of degrees of conventionality.

[33] For opposition to this, see Keefe (2000, ch.8); for a defence see **?**, p.123-5.

[34] Notice that the 'vagueness-related' supervaluationism should be distinguished from the 'inscrutability-related' supervaluationism of §3.1: these are two different deployments of the same theory. Essentially, the kind of supervaluationism we are

The natural thought is to look to *other* factors that counterbalance the difference in eligibility. For example, if supervaluational proposals are better than classical theories in predicative power or fit with the data, then the question of which is to be preferred is open. To evaluate the claim, we would need both a more developed account of the theoretical virtue of fit/predicative power,[35] and a way of reading off what the predications of classical theories/supervaluational theories are when filtered through an appropriate theory of competence and pragmatics. It is obvious, then, that the eligibility response offers no *easy* arguments for favouring classical over supervaluational semantic treatments of vague languages. However, if we worked through the details, and found that one overall package worked better than the other, I would not be unhappy to take that as a good reason for favouring the victor as the appropriate semantic treatment of vague language.

Objection 6: Reductive Presuppositions The eligibility response supposes that there are 'long chains of definition' for each property. Unless these definitional chains are finite, we will not have finitary theories to compare for syntactic complexity. The claim that *in general* such 'reductive' definitions are available seems extremely ambitious.[36]

Reply 6: Some of the hardest challenges here are really part of the problem of vagueness mentioned above. Finite definability of 'red' in terms of perfectly natural properties seems hard; finite definability of 'having surface spectral reflectance profile P' (which we may suppose to be a candidate for red-ness) is less so. We need only provide finite reductions for the salient candidate properties.

What is being presupposed is the disjunction: reductionism-or-emergentism. Wherever there are no finite chains of definability linking basic microphysics with a given property ('life' as it may be), we will need other perfectly natural properties that *do* allow for finite definability. In the context of an Armstrongian theory of universals, this amounts to 'emergent' sparse universals, of the kind that Armstrong allows for, but dislikes (Armstrong, 1978b, pp.69-71). What is ruled out is an attitude where properties 'merely supervene' on others, without being reducible in some stronger sense[37] In such scenarios, we would need to postulate emergent perfectly natural properties, or admit defeat.

The issues here are deep. 'Merely supervening' properties are something with which everyone should be uncomfortable: unexplained supervenience seems a metaphysical mystery. Without addressing this general question, there are cases which pose more direct difficulties. I have argued that one can give reductive paraphrases without being committed to reductive identifications or definitions. Theory shadowing paraphrases, such as that for 'reference' itself are a case in point. Such paraphrases were supposed to render supervenience on the properties involved in the base language unmysterious; but it is far from clear that they allow the possibility of finitary definition.

Of our six objections, the first three (inegalitarianism, the onion objection, and syntactic complexity) can be finessed without great cost. The latter three pose a more substantial challenge. In the case of optimality, one might adopt either a reductive or non-reductive realism; the core point being to resist subjectivism (resolving the problem, moreover, is something that is required by Humeanism about natural laws in any case). In the case of vagueness, there is at least a programmatic response.

here considering is one where the supervaluational theory is put forward as a *competitor* in the race for best theory. The supervaluational setting was used to 'encapsulate' all the successful (classical) candidate semantic theories to form a single overall meaning-fixing theory.

[35] Rothschild and Leuenberger (cite) suggest that the degree of fit of an interpretationist theory with its data might be based on a general account of 'truthlikeness' of theories.

[36] Sider (1995) raises this concern.

[37] Chalmers (1996) endorses such notions as a kind of reductivism.

The biggest worry by far is over the plausibility of finite definability. However, even this case is not clear-cut.

The situation is dissatisfying. The presumption built into Lewis' eligibility response that every property (or property-candidate) we refer to admits of finite reductive definition *seems* wildly overambitious. However, it is not clear how to turn this strong feeling into a focused objection to the proposal. In second half of this chapter, I will develop a more detailed objection to the eligibility response, based on inadequacy to its basic task: eliminating the threat posed by inscrutability arguments.

8.4 Eligibility and emergence

I want to outline a *prima facie* case that the eligibility response to inscrutability arguments is unsustainable unless a substantial metaphysical constraint is met. We already know that it is committed to property inegalitarianism. I shall argue that unless there are 'perfectly natural' properties at a relatively macroscopic level, then there are possible worlds macroscopically indistinguishable from our own, where an 'arithmetical' interpretation (where all singular terms denote numbers) is determinately *better* than the 'intended' interpretation (call such a situation 'Pythagorean'). Further, for all we know, our actual world could be Pythagorean—so we lose our claim to semantic knowledge.

I shall first briefly discuss the way of marking the divide between natural and non-natural properties in terms of which the discussion will be framed. I shall then point to a gap in the case against inscrutability: as direct rebuttal, the eligibility response is dialectically effective only against 'parasitic' arguments for radical inscrutability, and not against the completeness/compactness style arguments described in §5.4. Turning then to the global descriptivist form of interpretationism, I describe how to construct a finitary 'arithmetical interpretation' of the language, and use this to argue for the existence of Pythagorean worlds mentioned above. I finish by outlining the impact of this result, and the way that appeal to emergent universals (or equivalent machinery within other forms of property inegalitarianism) can resolve matters.

I frame this discussion in terms of a particular way of spelling out property inegalitarianism—a theory of sparse *universals*—although I think that the discussion can be reconstructed in the other frameworks also.

Universals

Armstrong (1978b, 1989) gives a vivid characterization of a certain kind of property ontology, that of *Universals*. These are to be objective existants, multiply located through space and time, incapable of existing uninstantiated, and wholly present at the location where they are instantiated. Such ontological characterizations as yet say nothing about the *distribution* of universals. A central part of Armstrong's view is that universals are relatively scarce. There may be a universal *being a Chimpanzee*, but there will be no universal *not being a Chimpanzee*; nor will there be a universal *being a Chimpanzee or being a Gorilla*.

Call a theory of universals *sparse* if by its lights there are no disjunctive universals, no negative universals and the like. A sparse theory then allows us to draw a distinction between the 'merely abundant' properties and the 'natural' ones, the latter being those holding of all and only the instantiators of some universal. All the theories of universals discussed here will be sparse in this sense.

Call a theory of universals *ultra-sparse* if the only universals that exist are those corresponding to the basic notions of microphysics: SPIN UP, CHARGE -1, and the rest.

Following Armstrong (1978b, pp. 69-71), I will call a universal *emergent* if it is a *simple* universal at a relatively high level: corresponding to the framework notions of chemistry, biology, psychology, for example. Emergent universals are to be distinguished, therefore, from *structural* universals, which within the Armstrongian theory, hold of relatively complex entities through (somehow) being 'composed of' more basic universals instantiated by the parts of the complex. *being H20* might be a structural universal; whereas if there is a universal *being an animal*, then it is likely to be an emergent.

A theory that incorporates emergent universals will be sparse, but not ultra-sparse. It may for example admit simple universals for chemical, biological and psychological kinds. Emergentism, for the purposes of this paper, is the thesis that there are universals for scientific kinds outwith microphysics. As such, it is rejected by both Armstrong (1978b) and by Lewis (1983a). It seems quite generally inconsistent with the position known as microphysicalism.

I want to emphasize three things about the framework here being used:

1. Emergence, as here used, is an entirely ontological matter. Often emergent properties are characterized epistemologically, in terms of their predictability from a 'lower-level' basis. That is not the case here.

2. I am neutral on the connections between arrangements of higher and lower level universals. There may be supervenience relationships between the distributions—and these might hold with metaphysical necessity, or the weaker nomic necessity.

3. I am neutral about some aspects of the nature of universals: whether or not they are 'cross-world' or rather 'world-bound' entities (Lewis, 1986c); and whether or not they have their nomic or causal roles essentially.

There are various reasons why one might wish to buy into emergent universals. One line of argument would focus on 'old work' for a theory of universals: their use in accounting for objective resemblance between things, in giving a framework for singular causation, and in providing a Realist account of laws of nature, and so forth (cf. Armstrong, 1978b, 1983, passim). If one wishes to make room for genuine macroscopic similarity, macroscopic causation, macroscopic laws and the like, under the universals-based analysis of these notions, it seems unlikely one will be able to do without emergent universals. Schaffer (2004) argues for emergence on exactly these grounds.

On the other hand, many find putative covariation of wholly distinct and metaphysical basic entities deeply mysterious. It is also hard to see how to avoid metaphysical vagueness, unless we restrict universals to the most fundamental microphysical kinds (cf. Barnes, 2005). As already noted, emergence contravenes the systematic reductive microphysicalism favoured by some.

As we shall see, the tenability of the eligibility response to inscrutability turns on this highly controversial metaphysical issue.[38]

A revenge problem for global descriptivism

Recall the way that eligibility allowed us to respond to permutation arguments for radical inscrutability. The appeal to eligibility relied on the parasitic nature of the permuted interpretations: for all we have said, it looks like the logical distance of *being the* ϕ*-image of something that is running* from the fundamental set of properties is greater than the logical distance of *being a runner* from that set. Our only grip on the twisted properties in terms of which the new interpretation is formulated, is derivative from the untwisted ones. We therefore have no reason to think that there is any way, in general, of characterizing these properties so as to match the most economical characterization of the original property in fundamental terms. There is a *prima facie* case that each twisted interpretation is less eligible than the standard interpretation. Dialectically, the argument for radical inscrutability fails.

Notice that there is no safety result: no guarantee that *every* permutation is such that on average the twisted properties are less eligible than the original. This is brought home by particular cases where permutation-style arguments plausibly succeed in the teeth of eligibility. Two cases where this can be argued are automorphic mathematical structures such as the complex numbers; and worlds containing structural symmetries (perhaps rotational or periodic).[39] Such limited gaps may not worry the friend of

[38] Though I shall present my discussion in terms of Universals, the distinctions just drawn, and the discussion below, can be replayed in terms of other ways of drawing the natural/non-natural distinction: trope theory, resemblance nominalism, or primitive naturalness, for example. As illustration, an ultra-sparse version of resemblance nominalism would have objective resemblance relations holding only between basic microphysical entities. Emergent Universals correspond to primitive relations of resemblance holding between relatively complex objects, allowing the picking out of natural properties at a relatively macroscopic level.

[39] Brandom (1996) and Strawson (1959) respectively, give versions of such arguments.

the eligibility response. Limited inscrutability of reference in these, very specialized, cases, would be far less shocking than radical inscrutability.

However, the situation changes if we use arguments for radical inscrutability that do not have the parasitic character of the permutation arguments. In particular, the completeness/compactness argument gives a *direct* construction of an interpretation matching the intended one at the level of truth-values or truth-conditions. Thus we have two interpretations, one built up out of a domain of arbitrary elements (we might as well take them to be numbers), and we have no grip on how the syntactic complexities of the two compare.

This should itself disturb the interpretationist. For the inscrutability argument at this point threatens to issue in stand-off—though the inscrutabilist has as yet no grounds for claiming her twisted theories match the intended theory on grounds of eligibility, neither does her opponent have any grounds for *denying* they do so. The stand-off is unhappy for the interpretationist, for the aim was to *secure* scrutable reference, not leave the matter undecided.

In fact, if we turn to the global descriptivist version of interpretationism, we can make the worry sharp.

The arithmetical interpretation

Suppose we adopt a global descriptivist form of interpretationism. The data constraining semantic theory in this case is a pairing of sentences with truth-values—or equivalently, a "total theory" that the semantics must render true. "Total theory" recall, was the sum total of all the platitudes gathered from every walk of life—all the sentences that are too obvious to question. If total theory is consistent, we know from elementary logic that we can find *models* for that discourse. Indeed, we have seen in detail in §5.4 how the completeness and compactness theorems of elementary metalogic give us recipes for constructing models in a systematic way, even where the language contains higher-order and intentional resources. For simplicity, however, let us suppose for the time being that total theory can be formulated in first order terms. We shall now use these resources to develop an argument for what I earlier called Pythagorean worlds.

We will be considering a case where this "total theory" is consistent with there being only finitely many things. For example, the total theory cannot commit itself to a plenitude of abstracta, or to infinitely divisible regions of substantival space or time. (If one holds that *our* folk theory of the world contains such commitments, consider a more cautious community who are at best agnostic about such matters). Take the consistent theory that results from *adding* "there are exactly n things" for an appropriate n, to total theory.

We can now appeal to our earlier discussion of the completeness based arguments to find a model whose domain consists of the numbers less than n. Call this the *arithmetical model* of total theory. Given this, it will be possible to write down long sentences listing element by element the objects assigned to a given predicate P. By brute force, then, we arrive at a description of the model. We can treat this as a semantic theory, whose clauses for predicates take an inelegant enumerative form.

What we now have available, in principle, is a semantics for total theory, which can be characterized in perfectly natural terms in some finitary 'enumerative' way. Since it is finite, we can simply read off the syntactical complexity of this theory. Say the syntactical complexity of the arithmetical interpretation is m.

Now we are in a position to see the danger for the global descriptivist. We have constructed an interpretation of the language whose elegance (when expressed in perfectly natural terms) is pegged at a particular natural number m. Moreover, we appealed to nothing about the world in constructing the model, beyond the fact that the particular total theory T was extracted. Hence, the same construction will be available in *any* world where T is total theory.

If the 'intended' interpretation for the language in question has greater complexity than *m*, when cashed out in perfectly natural terms, we're done—for both interpretations model global folk theory (i.e. fit the data) and in the scenario just mentioned, the intended interpretation will be less eligible than its rival, hence determinately unsuccessful. We may call worlds where the arithmetical interpretation beats the 'intended' interpretation in this way *Pythagorean worlds*.[40]

The existence of Pythagorean worlds

I shall now argue that, given the eligibility response to inscrutability and an ultra-sparse theory of universals of the kind that Lewis favours, there will be Pythagorean worlds.

In our world, let us suppose, atomic nuclei are made out of protons and neutrons, which are in turn made out of quarks. Only this bottom layer is a repository for universals, on the ultra-sparse conception. Presumably, that the microstructure of the world takes this shape is metaphysically contingent. Consider therefore a possible world that replicates ours from the 'quarks up', but in which the quark-counterparts are composed from yet more basic particles, the sub-quarks. The ultra-sparse conception now places the universals *below* the quark-counterparts.

We can iterate this procedure, each time adding underlying 'layers' to reality. On the ultra-sparse conception of universals, such underlayering pushes the domain of perfectly natural properties further and further away from the macroscopic. Moreover, each such iteration decreases the eligibility of properties such as *being Human*, since the long chains of definition down to the perfectly natural now need extra clauses added.

Notice, however, that this process allows us to decrease the eligibility of all of the properties which feature in the 'intended interpretation' *without limit*. For any number *N*, we can find a world like ours from the 'quarks up' with sufficient micro-structure that the syntactic complexity of the analogue of the intended interpretation, presented in fundamental terms, is more than *N*. In particular, we can choose a world of sufficient micro-structure that the syntactic complexity of the "intended" interpretation there exceeds the syntactic complexity *m* of the arithmetical interpretation. All the worlds we are discussing are like ours from the quarks up, so the same "total theory" will feature in each. Hence, such worlds will be Pythagorean.

(Note that even if we relax the restriction of the vocabulary of total theory, our case goes through. First, if we allow higher-order resources, the type theoretic constructions of Henkin (1950) (described in Appendix B) will allow us to make exactly the same case. If there are intentional resources present, we need to strengthen our assumption about total theory: we need to take a case where it is agnostic about whether there is an upper limit on the number of things that there could possibly be. In such settings, all possibility and tensed claims, etc, should be able to be rendered true while still allowing the kind of brute enumerative specification of a model that allowed us to argue for Pythagorean worlds.)

Three problems arising

The result is clearly disturbing. I see at least three direct problems.

First, and most obviously, we have a violation of an extremely intuitive principle that (some) semantic facts supervene on the macroscopic structure of the world. Worlds that are structurally like ours, but have different microphysical constituents might well engender twin-earth style differences in representational content, but there is no precedent for the kind of content change here envisaged. In particular, *A*-intensions or *linguistic meaning* of linguistic items (if one believes in such things) should supervene on macroscopic structure. In the Pythagorean world, *A*-intensions as well as 'horizontal content' (*C*-

[40] Compare Quine (1964).

intensions) are abandoned in favour of extensions drawn from the natural numbers.[41] Here we have a people behaving just as we do, within a world that is 'well-behaved' and just like ours in every detail from the quarks up; yet they refer to numbers where we refer to the objects around us. The case beggars belief.

Second, the actual world may itself be Pythagorean. The complexity of the arithmetical interpretation is enormous—but so are the lengths of definitions that would relate macroscopic structures to fundamental microphysics. I see no grounds for thinking that the first of these large numbers turns out, in the actual case, to be greater than the second.

Third, even if the actual world turns out to be non-Pythagorean, its being Pythagorean is a non-sceptical epistemic possibility. One route to this conclusion is just by noting the observation made above: that we have ourselves have no grounds for confidence that the intended interpretation will beat the arithmetical interpretation.[42]

This last conclusion is extremely disturbing. Part of our reason for rejecting radical inscrutability in the first place was to preserve epistemic access to semantic facts. We have just seen considerations to the conclusion that, even if semantic facts *do* happen to obtain, we have no reason to think that they do, and consequently no knowledge of any such facts.

Emergence

Notice the crucial role that the ultra-sparse conception of Universals played in the above argument. Suppose we abandon the ultra-sparse conception in favour of an 'emergentist' picture including macroscopic universals. On this picture, whatever chemical, biological, and psychological emergent universals we have are equally as 'fundamental' as microphysical kinds. Hence, the semantic theories that are compared with each other on grounds of syntactic complexity can be formulated in relatively macroscopic terms. The logical distance between the properties of which we speak and *these* properties is of course much smaller than that to the microphysical: correspondingly, there is little return in pressing worries that the distance might turn out to be greater than the astronomical complexity of the arithmetical interpretation. Moreover, since the elegance of the theories is now pinned relatively directly to macroscopic properties, one could not detract from the elegance of the theories by adding 'extra layers' to the base of reality.

[41] The point here is that the *whole semantic theory*—including the functions from contexts to horizontal content—is to be determined by best fit with appropriate data. In Pythagorean worlds, arithmetical extensions beat all such theories. Thanks to Bob Stalnaker for pressing me on this point.

[42] A more powerful, though more controversial route to this conclusion would be via the considerations some offer for holding that we have no justification for thinking that the actual world 'stops' at the level that current physics has detected (Schaffer, 2003). If this is right, the whole chain of worlds of increasing macro-complexities that we used to argue for the existence of Pythagorean worlds, are all epistemic possibilities.

8.5 Conclusion

The eligibility response to inscrutability response is a sound strategic line for an interpretationist to take. It is non *ad hoc*, given that it flows from a general analysis of theoretical elegance. It is dialectically effective against permutation arguments, even if there is no safety result.

A more general setting exposes serious limitations. Quite generally, when we consider arguments for radical inscrutability that are based on completeness and compactness, there are no longer direct grounds for disputing the inscrutabilist's argument. And, depending on how we cash out interpretationism, we may be able to argue *directly* for the existence of 'Pythagorean' worlds. I have argued that admitting the existence of such worlds is no better than admitting radical inscrutability.

We have a remedy, but it comes at a severe cost: the cost being what looks like an ontologically extravagant appeal to emergent property ontology (or its analogue within an ideological handling of the natural/non-natural distinction). The appeal would, of course, be less extravagant if one were independently committed to emergent universals and the like—perhaps in order to secure an analysis of macroscopic similarity, causation and laws. However, it is embarrassing, to say the least, that what looks to be a local problem for the metaphysics of meaning should require such substantial commitment.

We can expect a familiar dialectic to ensue. A *prima facie* case has been made for the need for a certain kind of ontology. One now looks around for something that can do the same work, at a cheaper philosophical cost.

One last direction for progress would be to revisit the way that Lewis handles theoretical elegance. What is needed to do the work is some objective way of picking out an appropriate vocabulary, using which we can compare theories with each other. Focusing on perfectly natural properties is a natural place to look. There may be other ways. For example, one might constrain the theories to be formulated in terms of intrinsic properties—and hope that this will rule out gerrymandered properties such as 'being such that T is true'. I will not speculate here on the prospects for this and related proposals that would attempt to pick out 'special' high-level properties without appeal to metaphysically primitive ontology or ideology.

What is clear is that some such underpinning is needed. Unless we patch the eligibility response in one of these ways, Lewis' remedy for inscrutability threatens to be worse than the disease.

Recapitulation and concluding remarks

Recapitulation

In the first part of this thesis, *The Framework*, I set out the basic challenges for a metasemantic theory. It is part of a more general project of giving a metaphysical account of representation properties: focused, in the present case, on the representational properties of language that are set out by semantic theories.

I proposed to investigate one particular strategy for addressing the metasemantic challenge. In interpretationism, unlike in causal theories, there is no suggestion of *identifying* semantic relations such as REFERENCE with relations that are in independent good-standing. Rather, (as described in Chapter 2) the "facts underlying" semantic statements are to be found in patterns of assent and dissent to sentences. §2.3 argued that this 'interpretationist' proposal is intimately related to fictionalisms, though it is to be divorced from any claim about the semantic *analysis* of the discourse in question. The key claim was that we can describe the facts underlying semantic talk by giving a 'theory-shadowing paraphrase', in terms of what follows from a *selected semantic theory*. The "real content" (in Yablo's sense) of semantic claims are those expressed by the fictionalist paraphrase.

Different implementations of this 'interpretationist' idea give different stories about how exactly these patterns should be described: for instance, Lewis identifies them with conventions of truthfulness and trust in uttering a sentence. Interpretationisms can differ also in how they take the data to constrain a 'selected' semantic theory. A common thread is that the meaning-fixing theory must be the *best* account of the relevant data—though here too, there is room for variation in what we take to be the theoretical virtues relevant to picking the 'best' theory.

In this setting, the possibility of *multiple* theories being selected became a worry. I presented two ways of handling such a situation: either construct an 'overall' meaning-fixing theory that embeds each of the others; or quantify over all the selected theories. In the latter case, any disagreement between the theories becomes a special case of the general puzzle of 'incomplete fictions'. I introduced the machinery in §2.2 for handling that puzzle; and in Chapter 3 compared it to the most promising "overall" account: the supervaluational framework. I concluded that there was not much to choose between these two options.

In general, where optimal semantic theories disagree with each other as to what it is that words (or other expressions) refer to, we get inscrutability. Some instances are innocuous: what I called 'framework' inscrutability, paradigmatically the choice of set-theoretic representation of relations. Other instances of inscrutability may illuminate puzzling features of language. Two potential examples are the case of theory-change described by Field (1973), and the problem of the many (Unger, 1980). It might be that vague language in general is a case of 'illuminating inscrutability' in this sense. Our focus then turned to two cases which are, at least at first glance, much more disturbing.

I argued in Chapter 4 that in the first of these cases—Quinean 'division' inscrutability—the initial appearance is misleading. First, we assessed the argument it was indeterminate whether 'gavagai' (or 'rabbit') divided its reference over rabbits (i.e. rabbit-worms), rabbit-stages or undetached rabbit-parts. There are challenging technical issues facing the Quinean case for this division-inscrutability. I suggested that the best rejoinder to technical worries was to lay out a semantic treatment of fragments of language, to enable a direct comparison of the results of 'dividing' reference in different ways. I concluded that, at least in the stage/worm case there is a strong argument for division inscrutability. Once we looked in detail at the candidate semantic theories, the appearance that we were setting bizarre re-interpretations against a common-sense view was shaken. Once described fairly, they all appeared equally bizarre—or better, all equally committal on matters where common sense had nothing to say. Moreover, none of the options had an obvious advantage on grounds of simplicity: they simply involve various divisions of labour between additional complexity in the form of extra primitives of the semantic theory (in particular, the counterpart relations invoked) vs. additional complexity encoded in the objects over which the quantifiers range and over which reference is divided.

The remaining case I consider is that of arguments for *radical inscrutability*. I developed permutation arguments for radical inscrutability in a highly general setting, which allowed us to see the potential power of the arguments. We also saw settings in which the arguments break down—though I maintained that if nothing further is said, radical inscrutability can still be argued for, since the underdetermination of selected theory by data is simply pushed back to indeterminacy in the data itself. We also found various generalizations of the inscrutability results: importantly, we have *radical indexical inscrutability*, which contends that there is no fact of the matter concerning whether a given term retains its reference from one moment to the next. This result was taken up in Chapter 7. An alternative 'direct' argument for radical inscrutability on the basis of Henkin's completeness/compactness proofs was described (§5.4), giving rise to the problems later discussed in §8.4.

Davidson thinks radical inscrutability of reference is innocuous, perhaps seeing it as a kind of 'framework' inscrutability of a kind with those described earlier. Rather than simply rely on incredulity at Davidson's position, I suggested we look for direct arguments against it. In the third part of the thesis, I investigate the theoretical basis of accepting radical inscrutability. I began by outlining a number of possible objections, but focused on two: potential tension between radical inscrutability and cognitive accounts of meaning; and potential tension between indexical inscrutability and inferential practice.

It is natural to think that inscrutability would 'infect' semantic beliefs—rendering beliefs about reference universally indeterminate. In Chapter 6, I argued that while the cognitive conception of reference is committed to lexical semantic beliefs (or something very like them), the infection principle itself is resistable given an appropriate account of the relationship between the content of language and the content of thought. We found no decisive case against inscrutability here.

In the case of inference, I gave considerations in support of Campbell's contention that "In our reasoning we depend on the stability of language, the fact that its signs do not arbitrarily change in meaning from moment to moment." (Campbell, 1994, p.82) I think that this principle is vital if we are to understand how facts about what follows from what can be relevant to token inferential moves. However, the indexical inscrutability result of §5.2 contradicts Campbell's principle. Here, I claim, is a real problem for the inscrutabilist.

The final part, and chapter, of my thesis focused on a way of avoiding inscrutability. I argued that we needed to look for a *principled* way of placing additional constraints on the selection of semantic theory—one that could be seen as deriving from theoretical virtues in general. Lewis' eligibility move, as I presented it, takes exactly this form. However, there is a revenge problem for the account: it relies on the parasitic nature of the permutation arguments, but has nothing dialectically effective to say about direct (compactness/completeness) arguments for radical inscrutability. In the case of global descriptivism, I showed how this led to the existence of 'Pythagorean worlds'—worlds where our terms (or their counterparts) denote abstract entities, though the world is macroscopically just like ours.

Their are patches available, but they come at a high cost: to directly patch Lewis' eligibility response we would need to endorse emergent universals at a macro-level. Perhaps the same job could be done at a lower cost: but that is a project we must leave for future work.

Concluding Remarks

Radical inscrutability arguments initially seemed to me to be a straightforward *reductio* of a simple-minded approach to what fixes meaning, easily avoided by more sophisticated accounts. My opinions have now changed, as the result of three factors. First, I am now more impressed with the difficulty facing other foundational theories of meaning. In particular, it is very difficult to see how metasemantics can deal with expressions other than paradigmatic terms for, and predicates applying to, medium-sized dry goods. Perhaps logical connectives and theoretical terms may be handled in a principled way: but what of 'very', 'of', 'the' and other ingredients of speech? What of the semantic significance of concatenation itself, reflecting in the compositional axioms of a semantic theory? Some sort of interpretationism seems to me, on reflection, highly attractive.

Second, when concentrating on a *reductive* account of the semantic, serious constraints are in force. The most straight-forward interpretationism gives a clear account of the way that patterns of assent and dissent determine a meaning-fixing theory: a theory is selected iff it fits the data. When we appeal to richer theoretical virtues, the account becomes much harder to formulate, particularly if we are concerned to avoid appealing to intentional resources. It is to Lewis' credit that he gives a picture of how this might work: but as detailed above, I think the theory is highly problematic.

Third, I am impressed with how difficult it is to make a compelling case for the *problem* with radical inscrutability. I considered a number of potential objections, but we needed to make serious commitments (concerning the nature of linguistic competence and the connection between inference and implication) before we had even a *prima facie* case. Even then, it was not clear that the inscrutabilist's resources are exhausted, as illustrated by the 'pseudo-semantic beliefs' offered to the inscrutabilist in §6.3.

I regard the theory of inference as posing the strongest objection to the inscrutabilist. If this could be finessed, the acceptance of radical inscrutability starts to look like a principled position. Part III of the thesis identified a number of resources, or 'substitutes' for the supposed explanatory role of scrutable reference.[43] The inscrutabilist should point out that he is not denying that there is a sense in which "Billy runs" is *about* Billy, rather than Jane (see §III and the introduction to part III); he merely insists that *reference* does not capture the sense of 'aboutness' in question. She might wish to avail herself of the pseudo-semantic beliefs of §6.3.[44] We can view such debate as investigating whether a key Davidsonian claim about the functional role of reference is correct; whether reference is:

> a theoretical construct, whose function is exhausted in stating the truth-conditions for sentences
>
> (Davidson, 1977, p.223)[45]

If inscrutability is accepted, reference-schemes would be presented as mere artifacts of systematizing data about truth-conditions, just as Davidson always insisted.

If the interpretationist is to accept inscrutability, however, she must be wary of deviant interpretations that have not been discussed here. A central worry will be the 'Kripkenstein' rule-following arguments (Kripke, 1980): a finite range of sentential data will always let us assign *truth-conditions* that are deviant,

[43] In the postscript to his (1978), Field suggests that non-semantic 'indication relations' may discharge some of the theoretic role of robust truth-conditional properties of sentences, leaving us free to deploy a disquotational treatment of semantic properties to discharge the residual theoretical role. Something like this is being suggested in the two cases below.

[44] Recall, also, that we did not close off the possibility that "pseudo-semantic beliefs" really do have a semantic subject-matter. We just noted that this would be a matter of controversy, and that there was no reason to require this in order for them to discharge the appropriate theoretical role.

[45] Page references are to the version collected in Davidson (1984).

so long as the sentences involved are not within the data-set. Such radical inscrutability of (certain) truth-conditions even less tenable than radical inscrutability of reference. Moreover, in the specific setting of global descriptivism, the original 'Putnam's paradox'—that whatever our (consistent) total theory is, our selected interpretation will render true—threatens the anti-realist result that total scientific theory is infallible.

Such puzzles may admit of resolution. Attention should turn to the form in which the sentential *data* is given. For example, if we appeal to *dispositions* to assent and dissent, rather than actualized patterns of such, we may be able to undermine the Kripkenstein considerations. Here is not the place to open up rule-following controversies, however.

Both for one concerned to resist inscrutability while avoiding complicating the meaning-fixing-theory/data relation, and one who wishes to defend radical inscrutability without falling victim to Kripkenstein and Putnamian worries, a focus on the *data* of interpretationism becomes important. Here we can no longer remain non-committal on details of how interpretationism is implemented: we must consider, in a case-by-case way, the virtues and vices of particular interpretationist proposals.

In my future work on this topic, the study of the form which the interpretationist *data* takes will be a principle focus. There are many different directions possible here: from appeal to fine-grained mental content, to causal connections, to structured states of affairs, to data constraining referential relations directly. On a 'head-first' approach, this will require attention to broader aspects of the problem of intentionality: particularly on the kind of structure we might be able to discern in perceptual and attitudinal content. An attraction of such approaches is that, like Lewis' eligibility response, they preserve the *principled* character of the interpretationist account of semantic facts.

In addition to this broad direction for tackling inscrutability, the work here presented has thrown up a number of challenging issues that would benefit from further work. I list some of the matters arising, chapter by chapter:

Chapter 1 As mentioned above, new and detailed forms of interpretationism may be needed in order to avoid inscrutability, incorporating new sorts of data, or more structured data of the traditional sentential kind. These may be based on one of the three models outlined in §1.3, or may take a different form. I am particularly keen to develop the deployment of decision theory sketched by Davidson (1980), and generally to examine how an interpretationist metasemantics fares in the language of thought setting described by Fodor (1987).

Chapter 2 One way of arriving at principled constraints on new interpretationist theories, is to think about the role of *underlying facts*. In Yablo's terms (Yablo, 2001), the real content of semantic statements can be formulated in terms of the patterns of assent and dissent. General reflection on the role of 'real content' of a discourse may therefore give us insight as to the acceptability, of, for example, causal word-object relations forming part of the interpretationist's data.

Chapter 3 A key question arising in Chapter 3 (also raised in §8.3), concerns the relationship between vagueness and inscrutability. If vagueness is to be understood as broadly a linguistic or representational phenomenon, rather than ontic or epistemic in character, then its relation to inscrutability becomes pressing. Is vagueness a form of inscrutability? Or do vague terms rather have special 'first-order' semantic properties, entirely distinct from vagueness? Even if vagueness is sometimes an ontic matter, how can a semantics be developed that reflects this? How can metasemantics cope with 'degree-vague' expressions such as 'red'? Similar questions can be asked about other forms of broadly indeterminate content, such as that generated by the problem of the many. The cases need to be handled with care, since the interaction between various views can be damaging.[46]

[46] See my Williams (2006a).

Chapter 4 A general project would look at the potential applications for dot theory: perhaps as a semantics compatible with mereological nihilism. In such cases, we might be willing to accept some of the quirks which make problems for division inscrutability.

Chapter 5 The extension of the permutation arguments in Appendix C to λ-categorial languages, while adequate for our purposes, would be more elegant if we had either a reduction of λ-categorial languages to pure categorial languages, or a fully general treatment of λ-terms.[47]

Chapter 7 As a self-standing treatment of the relation between logical consequence and good inference for an indexical language, one would hope to be able to be able to theorise about apparently good inferences that *exploit* context change: for example, the move from "*A* is *F*" uttered at one time, to "*A* was *F*" uttered at a later time. As noted on p.144, I do not think that it is *mandatory* to think of these as inferences supported by logically valid argument. Nevertheless, I think a setting can be developed where these inferences can be underpinned by a chain of valid arguments, within a purely model-theoretic setting. The key will be to transfer the ideas of Campbell (1994) and Prosser (2005) to the model-theoretic setting, and here I think the metaphysics of words developed in Kaplan (1990), (and endorsed by Prosser (op cit)) hold the key. I hope to develop this further in future work.

Chapter 8 In discussing Lewis' eligibility response by making a *prima facie* case for the need for emergent universals to avoid the threat of Pythagoreanism. I continue to think that Lewis' approach is the best motivated response to inscrutability considerations that focuses on esolving directly the underdetermination of semantic theory by the interpretationist's data. A general programme making the case for emergent universals, comparable to that offered in favour of property inegalitarianism in Lewis (1983a) would be one way of blunting the cost of the patch I offer the eligibility response. One approach would be to look at whether the original motivations for property inegalitarianism in Armstrong (1978b) and Lewis (1983a) require emergent properties; a step in this direction is made by Schaffer (2004). Alternatively, there is the programme of identifying 'special' higher-level properties without supposing robust property ontology that I mentioned at the end of chapter 8. We should not rule out, either, a quite different handling of the theoretical virtue of simplicity better able to discriminate between the complexity of the *subject-matter* of the theory and the lack of simplicity of the theory itself.

This thesis has not developed a metasemantic framework immune from inscrutability worries; neither has it brought comfort to those wishing to adopt a laissez-faire attitude to radical inscrutability. We have concentrated on diagnosing the source of inscrutability, understanding its nature and the nature of the framework that gives rise to it, mapping the space for responses to it, and examining the challenges faced in giving a principled response. The effect has been not to close down the original puzzles, but rather to sharpen them into a set of new and, hopefully, deeper challenges.

[47] As noted in Appendix C, the result there given is restricted to λ-operators applied to sentences (expressions of type *S*). The λ-terms of Cresswell (1973) had no such restriction.

Appendices

Appendix A

Is supervaluational consequence revisionary?

In the literature on supervaluationism, a central source of concern has been the acceptability or otherwise of its alleged logical revisionism. Timothy Williamson claims that supervaluationism gives rise to:

> . . . breakdowns of the classical rules of contraposition, conditional proof, argument by cases and *reductio ad adsurdum* in the supervaluationist logic of 'definitely'.
>
> Williamson (1994, pp.151-152)

Williamson is clearly unhappy with such revisionism:

> Conditional proof, argument by cases and reductio ad absurdum play a vital role in systems of natural deduction, the formal systems closest to our informal deductions.. . . Supervaluationists have naturally tried to use their semantic apparatus to explain other locutions. If their attempts succeed, our language will be riddled with counterexamples to the four rules.
>
> (ibid)

Others, accepting the case for logical revisionism, have argued that the upshot is unobjectionable. Thus Keefe:

> A number of commentators have emphasised how supervaluationist logic . . . fails to preserve certain rules of inference or classical principles about logical consequence
>
> . . .
>
> How important is the failure. . . of certain classical principles governing logical consequence? . . . My reply is that the described features of supervaluationism are acceptable. . .
>
> Keefe (2000, pp.176-178)

As both these authors emphasize, the case for logical revisionism depends on the presence, in the language at hand, of the supervaluationist notion 'Definitely' (the 'D' operator).[1] Roughly, 'Definitely p' says that, no matter how we sharpen the indeterminacy in our language, p always holds. In this, it is an object-language reflection of the supervaluationist's notion of truth—'supertruth': the idea being

[1] Though related notions, 'intentional' with respect to the parameter of delineations, would also lead to complaints.

that 'p' is supertrue if it is true no matter how we sharpen our language. In a supervaluational language without the D-operator and its relatives, there is no special threat to the classical modes of inference. Once it is added, it is alleged that the following results hold, providing counterexamples to the respective classical modes of inference.[2]

Contraposition

- $p \models_{SV} Dp$
- $\neg Dp \not\models_{SV} \neg p$

Conditional proof

- $p \models_{SV} Dp$
- $\not\models_{SV} p \supset Dp$

Argument by cases

- $p \models_{SV} Dp \vee D\neg p$
- $\neg p \models_{SV} Dp \vee D\neg p$
- $\not\models_{SV} Dp \vee D\neg p$

Reductio

- $p \wedge \neg Dp \models_{SV} \bot$
- $\not\models \neg(p \wedge \neg Dp)$

Here, for example, is Keefe on contraposition:

> in any specification-space where A is super-true, DA is also super-true since DA is defined as true whenever A is true on all specifications. However, it is not typically the case that $\neg DA \models_{SV} \neg A \ldots$ in a specification space where A is true on some specifications and false in others, $\neg DA$ is super-true while $\neg A$ is not.
>
> Keefe (2000, p.176)

Below, I argue that these results do not hold if the supervaluationist's framework is properly chosen. Moreover, even if they *did* hold, it is obscure in what sense they would count as a 'revision' of classical logic.

Responses

I distinguish two ways for the supervaluationist to respond to the arguments above. The first is to claim *contra* Williamson and Keefe, that supervaluational consequence is thoroughgoingly classical. The second is to accept that the examples arise and, with Keefe, to argue that the revisionism induced is not objectionable. It is the first line of response that concerns me here.

I can think of three ways of making the case that supervaluationism involves no departure from classical logic.

[2] The following are taken from Williamson (op cit)

1. In order to deny that the examples arise, one might give a non-standard treatment of the connectives involved, $\supset$, $\vee$ and so forth. One would make a case that, properly understood, $\models_{SV} p \supset Dp$ holds, so that the validity of $p \models_{SV} Dp$ does not lead to a failure of conditional proof.

2. One might try to undermine the case by characterizing $\models_{SV}$ in such a way that results such as $p \models_{SV} Dp$ does not hold, so that we can retain the standard treatment of connectives, and still not fall into revisionism.

3. One might make a case that the examples given above are not revisions of classical logic at all, because classical logic fails to sustain the relevant inferences. This would involve claiming that conditional proof, argument by cases and the rest are not universally valid, even for the thoroughgoing classicist.

I will focus here on (2) and (3). I argue that the natural generalization of the classical characterization of logical consequence will give a version of $\models_{SV}$ that does not lead to departures from classical logic: indeed, I will argue that none of the cases that Williamson and Keefe cite provide counterexamples to the rules they mention. Furthermore, I do this while accepting much of Williamson's setting: in particular, his rejection of 'local' characterizations of consequence in favour of 'global' characterizations.[3] In addition, I point to counterexamples to conditional proof, contraposition and the rest *within* classical logic.

The setting

Let me begin by outlining the treatment of consequence I favour. With Williamson (op cit) I characterize consequence model-theoretically. The first challenge is to say what a supervaluationist model looks like.

A supervaluationist model structure for a language L will consist, at minimum, of a domain of individuals D and a set of "delineations" Δ, and appropriate accessibility relations on that domain. In addition, there will be an interpretation function f, which will assign to expressions classical extensions relative to each delineation. Intuitively, at each delineation the function will project each expression onto a perfectly sharp 'meaning'. Formally, the model-structure so characterized is exactly analogous to that appropriate to a possible-worlds treatment of a modal language.

The crucial difference between my setting, and those of Williamson and Keefe, is that I assume that within the model structure of the intended model there will be delineations that are 'extreme'—relative to which a 6'8" man is short, for example. There are several reasons for wanting to have such delineations within our model structure. Consider the following attractive supervaluationist treatment of comparatives (cf. Lewis, 1970a; Kamp, 1975).

> 'A is F-er than B' is true iff the set of delineations where 'A is F' is true is a proper superset of the set of delineations where 'B is F' is true.

This will be untenable unless we have available extreme delineations. For otherwise, the set of delineations making-true 'A is tall' (where A is 6'8") will be the same as that making-true 'B is tall' (where B is 6'10"). On the treatment of comparatives given above, this will mean that 'B is taller than A' will be declared false, which is absurd.[4]

[3] Williamson gives arguments for the revisionary consequences. I do not dispute that the results follow from the framework for supervaluationism that he sets up—I dispute one element of that framework on which the cogency of his arguments turn.

[4] Williamson (1994) and Keefe (2000) cite such cases as objections to the treatment of comparatives, under their assumption that all delineations are non-extreme.

A second example: my desk is definitely flat. But in some extreme sense, it is not flat—it is less flat that an oil slick, for example. To give a treatment on which 'this is definitely flat, but in some extreme sense, it is not flat' will come out true, we

Within the intended model, then, we will find extreme delineations. Nevertheless, we want 'a 6'8" man is not short' to be true, and 'a 6'8" man is short' to be false, on the intended model. Due to the presence of extreme delineations, we cannot characterize supertruth in the simplest way, i.e. saying that S is supertrue iff on every delineation d, S is true relative to d. Some extra machinery is called for.

To finesse these issues, we first require models to pick out a subset of the delineations—a subset we will call the *sharpenings*. S will be supertrue (on a model) if it is true relative to each of the sharpenings of that model. A sentence will be supertrue *simpliciter* if it is supertrue at the *intended* model.

Second, we introduce an accessibility relation, S-access, on the space of delineations of the model structure, and let 'Dp' be true at a delineation d in a model if 'p' is true at all delineations S-accessed by d. On the intended model, the set of sharpenings will S-access each other.[5] Our models will therefore take the form:

$$m = \langle D_m, \Delta_m, S_m, f_m, s_m \rangle$$

where D_m is the domain, Δ_m the set of all (extreme and non-extreme) delineations, S_m an accessibility relation, and s_m a subset of the delineations—the sharpenings of the model.

Given this setting, we can then distinguish two forms of consequence. The first is *local* consequence:

> $\Gamma \models_{\text{local}} \phi$ iff
> On all models m, and all $d \in \Delta_m$, f_m makes Γ true relative to d only if f_m makes ϕ true relative to d

Williamson rejects this characterization, on behalf of supervaluationists. He takes it that consequence should be characterized in terms of *truth* preservation under arbitrary re-interpretations. For standard supervaluationists, then, it should be characterized in terms of supertruth-preservation. Given this, local validity looks suspect:

> The problem for supervaluationists is that supertruth plays no role in the definition of local validity. Yet they identify truth with supertruth; since validity is necessary preservation of truth, they should identify it with necessary preservation of supertruth. That amounts to an alternative definition...
>
> Williamson (1994, op cit)

This alternative is *global* consequence:[6]

> $\Gamma \models_{\text{global}} \phi$ iff
> On all models m, Γ is supertrue-at-m only if ϕ is supertrue-at-m

The notion of supertruth—truth at all sharpenings—is the central notion used to define supertruth on a model. Global consequence, as we have characterized it, meets Williamson's constraints. I shall assume in what follows that $\models_{\text{global}}$ is a proper explication of $\models_{\text{SV}}$.

need to appeal to extreme delineations 'accessed' by 'in some extreme sense'.

Yet another reason for wanting such delineations is the need to treat higher-order vagueness within the supervaluationist setting. See Williamson (1994, §5.7).

[5] Another constraint on S will be, presumably, that it is reflexive. More constraints will presumably flow from an adequate account of higher-order vagueness.

[6] Williamson does not mention the relativization of truth to a model in the above. However, I take it that he will wish to generalize over models in the final characterization, on pain of admitting water=H_20 as a logical validity.

If we added into our models a set of possible worlds, and a specification of one among these as 'actual', then generalizing over all models would amount to requiring *necessary* truth preservation, as Williamson requires.

Incidentally I am not personally committed to rejecting local characterizations of consequence, but shall not argue the point with Williamson here. Since I have no *objection* to the global characterization, I am independently interested in exploring its behaviour.

Revisionism?

Does $\models_{\text{global}}$ induce logical revisionism? I argue not. Consider, for example, the alleged result that $p \models_{\text{SV}} Dp$. When $\models_{SV}$ is read as $\models_{\text{global}}$, then we can find counterinstances. Take a model where the set of sharpenings contains a single delineation δ, relative to which p is true (thus p is supertrue at that model). Assume further that the accessibility relation S relates δ to a delineation δ', and p is false relative to δ'. Then 'Dp' is not true at δ, and hence not supertrue-at-m. Hence, we have a countermodel to the claim that Dp is a consequence of p. The upshot is that this sequent cannot play a role in showing that contraposition, or conditional proof, fails.

Countermodels can be found to the other examples too. For example, consider the alleged result that $p \wedge \neg Dp \models_{SV} A \wedge \neg A$.[7] The model above is a counterinstance to this also; for relative to the model just described, the premiss is supertrue, but the conclusion superfalse. Indeed, all the results cited fall to such considerations.

Notice that to find these countermodels, the S-accessiblity relation has to relate sharpenings to delineations that are not themselves sharpenings. One might think this illegitimate: the accessibility relation was introduced precisely to enable a definition of 'D' that *reflected* the set of sharpenings.

I claim this misunderstands semantic consequence. The basic idea of that notion is that ϕ will be a semantic consequence of Γ iff whenever the latter is true, the former will be true also *no matter what the expressions involved may mean*. That the above model gives an *unintended* interpretation of 'Definitely'—i.e. so that it does not reflect supertruth—is no objection to our citing it as a counterinstance to a claimed consequence relation. An unintended interpretation is exactly what one should look for, in general, to find counterexamples to claims that one thing is a semantic consequence of another.[8]

Now of course *total* re-interpretation of the particles of language won't give us a classical notion of consequence. For example, if we are allowed to re-interpret 'and' as meaning *or*, then $A \wedge B \models A$ will have counterinstances. Accordingly, the model-theoretic treatment allows us to declare *inadmissible* certain models; paradigmatically, those where the interpretation of logical constants deviates from that intended. In the case of modal logic, we standardly declare models inadmissible when the accessibility relations do not obey certain formal features of the accessibility relation we wish to preserve—which features these are depends on the modality we are interested in.

One aiming to convict supervaluationism of revisionism may say that the models we cited above are similarly inadmissible. But what could possibly be the motivation for this? Not any analogy with precedent within classical modal logic: normally accessibility relations are allowed to vary as they might, subject only to *formal* requirements (e.g. reflexivity, as it might be). Not from any desire to preserve classical logic more generally, since ruling out such models will mean *giving up* on conditional proof, contraposition and the rest. Not from a need to secure other less controversial inferences: for example, to secure that p will be a consequence of Dp, we need only insist that in every logically admissible model the S-access relation is reflexive—exactly the kind of restrictions on admissible models familiar from modal logics more generally.

The only way of arguing that our counterinstance should be ruled inadmissible, as far as I can see, is a direct approach: one would claim that the connection between the accessibility relation used to define D, and the set of sharpenings that a model provides, should be taken as 'logical'. Again, the question is: what is the argument for this?[9] Certainly, the Tarskian characterization of logical constants, via

[7] Williamson (1994, fn.5.19 p.297) cites this as a robust result, one from which counterexamples to the above modes of inference all follow.

[8] Consider, for example, a counterinstance to a claim that 'Hesperus=Phosphorus' is semantically valid. This will have to involve assigning to either 'Hesperus' or to 'Phosphorus' something that it does not designate—that is, we appeal to an unintended interpretation of this bit of language.

[9] In conversation, Williamson has indicated that he does regard DEF as a logical constant in these circumstances: and indeed, he would favour treating modal operators such as metaphysical necessity in parallel ways, rather than following the

permutations, is unlikely to deliver such a result.[10]

I conclude that we should let the S-access relation vary within logically admissible models independently of which delineations count as sharpenings in that model. Accordingly, the argument that global definitions of consequence entail logical revisionism fails.

Counterexamples to the rules without delineations

We can find counterinstances to contraposition, conditional proof, argument by cases and reductio within a purely classical setting. This is due to an orthodox treatment of variables, rather than the novel treatment of delineations. Standard model-theory for a predicate logic works with a notion of truth relative to a variable assignment. For example, 'x is male' will be true relative to an assignment that pairs x with Tony Blair; and false with respect to an assignment that pairs x with Margaret Thatcher. Standardly, we say that x is true (/false) on a model iff it is true (/false) with respect to *every* variable assignment.[11] Immediately, we find 'failures of bivalence' involving open sentences. It follows directly from the definitions and observations made above that, on the intended model, "x is male" is neither true nor false.

The considerations in favour of global consequence that Williamson offers transfer directly to this case: consequence, as truth-preservation, should be defined in terms of truth-on-a-model. Given this, it is simple to verify that we have the following counterexamples to "classical" rules:

Contraposition

- x is male $\models \forall x(x$ is male$)$
- $\neg\forall x(x$ is male$)\not\models \neg(x$ is male$)$

Conditional proof

- x is male $\models \forall x(x$ is male$)$
- $\not\models x$ is male $\supset \forall x(x$ is male$)$

Argument by cases

- x is male $\models \forall x(x$ is male$)\vee\forall x\neg(x$ is male$)$
- $\neg x$ is male $\models \forall x(x$ is male$)\vee\forall x\neg(x$ is male$)$
- $\not\models \forall x(x$ is male$)\vee\forall x\neg(x$ is male$)$

Reductio

- x is male $\wedge\neg\forall x(x$ is male$) \models \bot$
- $\not\models \neg(x$ is male $\wedge\neg\forall x(x$ is male$))$

standard treatment of modal consequence where domains and accessibility relations are allowed to vary. Clearly, these are controversial issues; and substantial support for this approach would be needed if the view is to be dialectically effective against the supervaluationist.

[10] See MacFarlane (2000) on extending the Tarskian permutation criterion to the intensional case. He argues that it is hard to see how such a categorization could justify restricting the accessibility relation in the formal ways described above. If so, this makes our case even stronger.

[11] Lewis (1970a) suggests we take the analogy to its fullest extent and think of variable-assignments as indices, alongside worlds, times, delineations etc. within a multiply intensional general semantics.

We have here a definition of truth-on-a-model, characterized by generalizing over all indices of a certain kind—in this case variables. We also have a certain operator—universal quantification—which reflects this definition of truth, and indeed, is constrained to do so on all models. Williamson finds analogous behaviour within the setting he offers the supervaluationist because supertruth and 'definitely' stand in an analogous relationship, for him. For Williamson, a sentence is supertrue if true on all delineations; and 'Dp' is true on an arbitrary delineation in the same circumstances. In the treatment of supervaluationism that I have been offering, this connection between the definition of truth and the definition of an operator is broken; so the phenomena do not arise.

If required, we could perform a similar trick with the above results. We could characterize the universal quantifier in terms of an accessibility relation over the variable assignments, which was then allowed to vary in unintended models.[12] Two factors are missing that were present in the supervaluational case: in that case we had motivation for formulating the 'D' operator in terms of an accessibility relation, once we had (for independently motivated reasons) introduced extreme delineations into our semantics. Second, there was no *prima facie* case for regarding D as a logical operator, whereas there is such a case for the universal quantifier. The case for "revisionism", if such it is, is much stronger in the case of free variable formulas, than it is in the case of 'Definitely'.

Conclusion

One headache for supervaluationists in recent times has been the logical revisionism allegedly implicated by their setup. I say that there is a natural framework for supervaluationism that undermines the arguments for revisionism offered in the literature. Moreover, the setup—involving extreme delineations—is one that supervaluationists have independent reasons to accept.

Perhaps the case for revisionism can be mended: one could try to argue that there are good reasons for taking the link between 'Definitely' and the set of sharpenings to be 'logical' in nature, contrary to what has been urged above. It is likely that such a case would involve deep and controversial claims about the nature of logicality, particularly in relation to sentential operators such as 'definitely', 'necessarily', 'possibility' and even quantifiers. The burden of proof is certainly on the supervaluationist's critic to make such a case.

Our final observation is that there are counterexamples to all the rules cited earlier within a purely classical setting. This makes urgent the need for clarification, from the supervaluationist's critic, on what counts as an objectionable 'logical revision'. We already have precedent for the kind of breakdowns in contraposition, conditional proof and the rest that Williamson cites. Further, reasoning with open sentences does take place—in mathematics at least. Again, the burden is on the critic to make the case that any *supervaluation*-related failures are specially objectionable.

[12] To get this to work, we may need to allow it to access individuals from *outside* the domain.

Appendix B

Completeness and compactness

The results and proofs below are based on those given in Zilber (2000), and the ideas derive ultimately from Henkin (1949).

We first consider a first-order language, $\mathcal{L}$, which is constructed in the usual way from logical symbols (including identity, '$\doteq$') and a non-logical vocabulary of constants, function-symbols and predicates. We will consider *arbitrary* such languages, imposing no limitations on the cardinality of the language.

Theorem 2 (Completeness)**.** *Any consistent set Σ of $\mathcal{L}$-sentences is satisfiable. (In particular, it is satisfied by a structure of cardinality no greater than that of the language.). Equivalently: if a set of sentences is unsatisfiable (i.e. $\Gamma \Vdash \bot$) then it is inconsistent (i.e. $\Gamma \vdash \bot$).*

A corollary is:

Theorem 3 (Compactness)**.** *Any f.s. set Σ of $\mathcal{L}$-sentences is satisfiable. (In particular, it is satisfied by a structure of cardinality no greater than that of the language.)*

This follows from the observation that a set is finitely satisfiable only if it is consistent. Consider an inconsistent set Λ. There must be a finite derivation of an absurdity from this set—the derivation will appeal to only finitely many elements of Λ. Conjoining, and appealing to the deduction theorem, we have $(\lambda_1 \wedge \ldots \wedge \lambda_n) \to \bot$ derivable from the null set. Since $\bot$ is not satisfiable, by the interpretation of $\to$ every model must make one of λ_i false. Hence this finite subset of Λ is not satisfiable.

We shall prove the completeness theorem and allow compactness to fall out as a corollary. The proof given could easily be modified to prove compactness directly, substituting 'finite satisfiability' for 'consistency' throughout, and making the necessary small changes.

In the following, we will need the following notions. A set of $\mathcal{L}$ sentences Σ is *maximal* if, for every $\mathcal{L}$-sentence ϕ, either $\phi \in \Sigma$ or $\neg\phi \in \Sigma$. It is *full* if whenever $\exists x\phi(x) \in \Sigma$, there is some constant c such that $\phi(c) \in \Sigma$. It is finitely satisfiable if every finite subset $\Sigma_0 \subseteq \Sigma$ has a model. It is *deductively closed* with respect to finite subsets if, whenever Γ is finite subset of Σ, and $\Gamma \models \chi$, then $\chi \in \Sigma$. A *canonical* model for Σ is one in which every element of the domain is denoted by a constant symbol in the language.

We now move to show that completeness holds for first order languages with arbitrary sets of non-logical vocabulary.

Lemma 4. *If Σ is maximal and consistent then it is deductively closed with respect to finite subsets of formulae.*

Proof. Suppose $\phi, \psi \in \Sigma$ and that $\phi, \psi \vdash \chi$ ($\Leftrightarrow \{\phi, \psi, \neg\chi\}$ is inconsistent.) By maximality of Σ, either χ or $\neg\chi \in \Sigma$. If $\neg\chi$ were in Σ, then $\{\phi, \psi, \neg\chi\}$ would be a clearly inconsistent subset of Σ. So $\chi \in \Sigma$. This reasoning generalises to arbitrary finite cases.[1]

■

Proposition 5. *For any full and deductively closed set Σ of $\mathcal{L}$-sentences there is a canonical model $\mathcal{A}$ of Σ. Further, this canonical model can be chosen to embed an arbitrary reference-scheme σ, so long as this is consistent with the identities contained in Σ.*

Proof. Note that it follows from the above that Σ is deductively closed wrt finite subsets.

Let Λ be the set of closed terms of $\mathcal{L}$. For $\alpha, \beta \in \Lambda$ define an equivalence relation $\sim$ that holds iff $\alpha \doteq \beta \in \Sigma$. That this is an equivalence relation is established by the above remark and the elementary properties of identity.

Let A be the set of equivalence classes of Λ under $\sim$. The domain of our canonical model could be taken to be A. However, we can build a canonical model from any domain of objects that is bijective with A. Suppose, in particular, that we wish to construct a canonical model that matches some function σ from terms of $\mathcal{L}$ to objects drawn from O. Suppose also that $\sigma(\alpha) = \sigma(\beta)$ whenever $\alpha \sim \beta$.

Extend take any superset M of O, of cardinality equal to A, and pick some particular bijective function $\phi : A \approx M$. Now write $\tilde{\alpha}$ for the image under ϕ of the equivalence class of α under $\sim$.

We need now to assign an interpretation of constants, functions and relations of $\mathcal{L}$ relative to M. This will proceed in the obvious way. Thus:

- $c^{\mathcal{A}} = \tilde{c}$
- $f^{\mathcal{A}}(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n) = f(\widetilde{\alpha_1, \ldots, \alpha_n})$
- $\langle \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n \rangle \in R^{\mathcal{A}} \longleftrightarrow R(\alpha_1, \ldots, \alpha_n) \in \Sigma$

It is easily checked that these definitions do not depend on the choice of representatives for $\tilde{\alpha}$. If α_i and β_i are different representatives of the same equivalence class, then we have $\alpha_1 \doteq \beta_1, \ldots, \alpha_n \doteq \beta_n \in \Sigma$. Note that: $\alpha_1 \doteq \beta_1, \ldots, \alpha_n \doteq \beta_n, P(\alpha_1, \ldots, \alpha_n) \models P(\beta_1, \ldots, \beta_n)$. By deductive closure of Σ, it follows that $P(\alpha_1, \ldots, \alpha_n) \in \Sigma$ entails $P(\beta_1, \ldots, \beta_n) \in \Sigma$. Similar arguments show the well-definedness of other clauses.

This completes the definition of the $\mathcal{L}$-structure $\mathcal{A}$ which is to be the canonical model. We may now show that it does indeed model Σ. It suffices that:

$$\mathcal{A} \models \phi(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n) \Longleftrightarrow \phi(\alpha_1, \ldots, \alpha_n) \in \Sigma$$

We proceed by induction on the complexity of ϕ.

The base cases hold by construction. For propositional connectives, the deductive closure of Σ gives our result. In the case of quantifiers, we need to use the *fullness* of Σ. Let us suppose that $\exists x \phi(\alpha_1, \ldots, \alpha_n, x) \in \Sigma$. Now by the definition of satisfaction, $\mathcal{A} \models \exists x \phi(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n, x)$ holds if and only if there is some $\tilde{\beta} \in \mathcal{A}$ such that $\mathcal{A} \models \phi(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n, \tilde{\beta})$. By construction of $\mathcal{A}$, this last holds iff $\phi(\alpha_1, \ldots, \alpha_n, \beta) \in \Sigma$. This is the case just when $\exists x \phi(\alpha_1, \ldots, \alpha_n, x) \in \Sigma$ ($\Rightarrow$ by deductive closure, $\Leftarrow$ by fullness of Σ). This gives our result, once we note that the domain of $\mathcal{A}$ must have cardinality $\leq |\mathcal{L}|$, by construction.

■

The above result gives rise to the following general strategy for constructing models for sets of formulae:

[1] NB: notice that parallel reasoning would show that maximal finitely satisfiable sets are deductively closed with respect to finite sets.

> Take a consistent set of $\mathcal{L}$-sentences Σ and extend them to a set $\Sigma^\dagger$ in a language $\mathcal{L}^\dagger$ which is maximal consistent and full. By the first two conditions, it is deductively complete, so we can use the above construction to build a canonical model. We can then restrict the result to obtain an $\mathcal{L}$-structure modelling Σ.

A first step is the following lemma.

Lemma 6 (Lindenbaum Theorem). *Given a consistent set of $\mathcal{L}$-sentences Σ, there is a maximal consistent set of $\mathcal{L}$-sentences Σ^* extending Σ.*

Proof. By the axiom of choice, the set of formulae of $\mathcal{L}$ can be well-ordered. Since the language is set-sized, under this ordering it is order-isomorphic to some ordinal β. Index each formula by ordinals $\leq \beta$.

Take a set of $\mathcal{L}$-formula Σ, and perform the following construction:

$$\Sigma_0 \quad := \Sigma$$

$$\Sigma_\lambda \quad := \bigcup_{\gamma<\lambda} \Sigma_\gamma \qquad \lambda \text{ a limit ordinal}$$

$$\Sigma_{\alpha+1} \quad := \begin{cases} \Sigma_\alpha & : \quad \text{if there is no } \phi_\alpha \text{ in } \mathcal{L} \\ \Sigma_\alpha \cup \{\phi_\alpha\} & : \quad \text{if this is consistent} \\ \Sigma_\alpha \cup \{\neg\phi_\alpha\} & : \quad \text{otherwise} \end{cases}$$

Set $\Sigma^* := \Sigma_\beta$ (note that $\Sigma_\gamma = \Sigma_\beta$ for all $\gamma > \beta$)

Note that, so long as Σ_α is consistent, one of $\Sigma_\alpha \cup \{\phi_\alpha\}$ and $\Sigma_\alpha \cup \{\neg\phi_\alpha\}$ must be consistent. We can see this by *reductio*. If both were inconsistent then $\Sigma_\alpha \cup \{\phi_\alpha\} \vdash \bot$ and $\Sigma_\alpha \cup \{\neg\phi_\alpha\} \vdash \bot$. Appealing to the deduction theorem twice, $\Sigma_\alpha \vdash \phi_\alpha \rightarrow \bot$ and $\Sigma_\alpha \vdash \neg\phi_\alpha \rightarrow \bot$. Logic then gives us that: $\Sigma_\alpha \vdash \neg\neg\phi_\alpha$ and $\Sigma_\alpha \vdash \neg\phi_\alpha$, and so $\Sigma_\alpha \vdash \neg\neg\phi_\alpha \wedge \neg\phi_\alpha$. So Σ_α itself would be inconsistent.

With this in place, it is easy to see that the construction preserves consistency. Given that the initial Σ is consistent, Σ^* will be a consistent set.

Σ^* will also be complete. Take an $\mathcal{L}$-sentence ψ. This has a place in the well-ordering described above, so $\psi = \phi_\alpha$ for some $\alpha < \beta$, and so one of ψ, $\neg\psi$ will be in $\Sigma_{\alpha+1} \subseteq \Sigma^*$.

■

Note that even without the axiom of choice, the above theorem would be available for languages with well-orderable sets of non-logical terminology. In particular, it would hold for countable languages.

Theorem 7 (Completeness). *Any consistent set Σ of $\mathcal{L}$-sentences is satisfiable. (In particular, it is satisfied by a structure of cardinality no greater than that of the language.)*

Proof. Extend $\mathcal{L}$ to $\mathcal{L}^+$ by adding constant symbols c_ϕ for each $\mathcal{L}$-formula $\phi(x)$ with exactly one free variable. Now form Σ^+ by adding the $\mathcal{L}^+$-formulas of the form "$\exists x\phi(x) \rightarrow \phi(c_\phi)$" to Σ. Notice now that Σ^+ is now *full* in $\mathcal{L}^+$.

It will be also be consistent if Σ is, for we would be able to replicate any proof of a contradiction from Σ^+ within the original setting, by appealing to existential elimination on appropriate instances of $\exists x\phi(x)$.

We now take the full, consistent set Σ^+ and use the Lindenbaum theorem to find a maximal superset Σ^{+*}.

This new set may not be full. However, by iterating the above procedures we get a chain of sets of formulae:

$$\overbrace{\Sigma}^{0} \subseteq \overbrace{\Sigma^+}^{1} \subseteq \overbrace{\Sigma^{+*}}^{2} \subseteq \overbrace{\Sigma^{+*+}}^{3} \subseteq \overbrace{\Sigma^{+*+*}}^{4} \subseteq \cdots$$

The non-zero even elements of the chain are full, the odd elements are maximal, and all are consistent. The union of the chain, Σ^{∞} will then be a maximal consistent and full set of sentences of the language $\mathcal{L}^{\infty} = \bigcup_{n\in\mathbb{N}} \mathcal{L}^{(+)^n}$.[2] By an earlier result, we will be able to construct a canonical $\mathcal{L}^{\infty}$-model $\mathcal{A}^{\infty}$ for Σ^{∞}. Note that none of the extensions of the language increase its cardinality, and hence this model has cardinality $\leq |\mathcal{L}|$. By appropriate elimination of the interpretation of new constant-symbols from this structure, we will obtain an $\mathcal{L}$-structure which will model the $\mathcal{L}$-sentences within Σ^{∞}, which is exactly Σ. ■

This proof depends on the axiom of choice insofar as it relies upon the Lindenbaum lemma. We may note, therefore, that the result can be proved without use of the axiom of choice in the case of well-orderable languages, and as a special case of this, for countable languages.

Given the proof just sketched (and in particular, the closely related version where 'consistency' is replaced by 'finite satisfiability' throughout), it is easy to go on to prove a version of the upward Loweinheim-Skolem theorem. Details can be found in Zilber (2000).

Extending the results to type theories

Completeness and compactness results can be proved for second order logics, and type theories, provided that we do not allow models which are not 'standard'. The 'non-standardness' of such models is not worrying for our purposes: it simply means that we allow the *domains* of the various types to vary in certain ways. From one perspective, this is a natural extension of the standard feature of first-order logic: that the domain of the first-order quantifiers should be allowed to vary from model to model. It may well be that the *intended* model of a second order or type theoretic language has its domains related to each other in a tightly constrained way: but this can be admitted while allowing admissible (unintended) models where this is not the case. The more general kind of model are known as 'general' or 'Henkin' models, as opposed to 'standard' or 'full' models.

There are two ways of viewing the extension of compactness results to the higher order cases. The first is to view a higher-order logic, say with 'second-order' (predicate-position) quantifiers, as a *multi-sorted first order language*. Predication would be represented by a relation holding between terms of the first sort, and terms of the second sort. 'Fx' is represented Ⓟ(F,x) (for the full type theory, we extend this to include many more sorts). We then need only generalize the above techniques to apply to relevant kinds of multi-sorted logics, and we can derive our result. See Shapiro (1991, SS4.3) for a full proof of compactness for second-order logic that takes this kind of line.

Henkin (1950) gives a proof of compactness for an extensional type theory that does not require such re-interpretation. The basic idea is a generalization of that given above. Given a consistent set Λ, one extends it via Lindenbaum-style techniques to a complete set Γ. (Within a Church-style type theory setting, this will automatically be full, as the axioms governing the λ and ι operators within the framework ensure that the maximal consistent set of sentences Γ will contain a singular term $\iota(\lambda x Fx)$ witnessing Fx whenever $\exists x Fx$ is in Γ.)

Henkin then puts closed expressions of each type into equivalence classes relative to Γ. Introducing the notion of identity '$=$' between expressions of a given type via a version of Leibniz's law, he lets expressions A_c, B_c be equivalent iff $\Gamma \vdash A_c = B_c$.

With this in place, he begins to build the domains. As before, the 'first order domain' (i.e. the domain of category N) will comprise equivalence classes of individuals. The category of S is given the domain $\{T,F\}$. In general, he argues that by suitably choosing the domains of each derived category, one can set

[2] Consistency: any finite subset of Σ^{∞} will be also a finite subset of some Σ_n, which is consistent. Since a set is consistent iff its finite subsets are consistent, this gives the result. Full: any $\exists x\phi(x)$ will be found in some Σ_n, and hence a witnessing constant will be introduced in the next constructions. Completeness: Any formula ϕ of $\mathcal{L}^{\infty}$ is a formula of some $\mathcal{L}^{(+)^n}$. By construction, either ϕ or $\neg\phi$ will be entailed by Σ_{n+1}, which is a subtheory of Σ^{∞}.

up a 1-1 function Φ from elements of the domain of category Q to equivalence classes of expressions of that category. This clearly holds for the first two cases. The equivalence classes of S are just two: those proved by Γ which we map by Φ to T; and the rest, which are mapped to F. The equivalence classes of N just are the elements of that domain, so Φ can here be taken as identity.

For a derived category $Q = R/T$ we set up an appropriate domain (by induction) as follows. We suppose that we have extended Φ to expressions of category R and T. An element of the domain will be a function, from elements of the domain of T to elements of the domain of R. So take q of category Q, and an arbitrary element A in T. By our induction hypothesis, there is some expression t of category T such that Φ maps the equivalence class of t to A; $\Phi([t]) = A$. Now consider $q(t)$, which is an expression of category R. By induction again, there is a unique corresponding element of R, B, such that $\Phi([q(t)]) = B$. Hence we set:

$$\Phi([q]) : A \mapsto B$$

It is easy to check that our construction depends at no point on the choices of representatives, and that we have thereby proved our induction step.

Henkin then proves a lemma which shows that, if we let variable assignments range over the frame of domains just characterized, we have a valuation of the language which matches Φ on the value it assigns to closed formulae. We *started out* by ensuring that Φ paired elements of Γ (and a fortiori, those of Λ) with T. Given this is indeed embedded within a valuation of the language as a whole, there is a valuation which makes each element of Λ true. (Furthermore, because each of the (countably many) domains is in 1-1 correspondence with a set of equivalence classes of expressions of the (countable) language, the overall domain is at most countable.) This provides the desired completeness result: for every consistent set of closed sentences, there is a valuation on which every member of the set is true.

Compactness is then a trivial corollary: established by linking finite satisfiability to provability as above.

Appendix C

Variables, abstraction and inscrutability

The pure categorial framework described in §5.1 has difficulties in giving an elegant treatment of quantification for non-basic categories. Consider the following two expressions: "For everyone, there's someone who loves them" and "For everyone, there's someone that they love". The most elegant approach would be to formalize these using λ-abstracts:

- Everyone λy(Someone(λx[x loves y]))
- Everyone λx(Someone(λy[x loves y]))

What Cresswell (1973) calls a 'λ-categorial' language treats λ as a new primitive—a syncategorematic expression, whose interpretation is not allowed to vary across models.

In order to keep their categorial languages 'pure', Lewis (1970a) and Montague (1970) treat λx for an appropriate variable x as a part of the lexicon. Lewis calls them 'binders'. He considers only the case where the variable is of category n, and treats λx itself as having category $(S/N)/S$. The key here is that Lewis treats variable assignments themselves as an index within the general semantics—this then allows intensions for the binders to be defined. The details can be found in Lewis (1970a, p.210-212).[1]

There is a serious limitation to Lewis' approach, which he notes in the postscript to (1970a). As Lewis sets it up, a variable assignment is a function from variable-numbers[2] to extensions of the appropriate type (in the case of category N variables, it is a function from variable-numbers to objects). If semantic values for all expressions were 'Carnapian' in form—i.e. functions from indices to extensions, we would have little problem in generalizing Lewis' idea. A variable assignment would simply map variables of category c to extensions of the type appropriate to c.

However, on Lewis' treatment, the semantic values of derived categories are 'compositional' rather than 'Carnapian'. The semantic value of an expression of category S/N, for example, will be a function from intensions of type $\langle i,e\rangle$ to intensions of type $\langle i,t\rangle$. Therefore, in the general case, there simply are no 'extensions' to range over. Moreover, we cannot generalize Lewis' idea by taking variable assignments as mappings from variables to appropriate *intensions*. Considered set theoretically, the indices themselves (in particular, all variable assignments) are in the transitive closure of any intension. If a

[1] Here is a quick sketch. We want to assign an intension appropriate to the category $(S/N)/S$ to λx_n. This will be a mapping from sentence intensions to (a function which maps name-intensions to sentence-intensions). Suppose a sentence intension f is input. We now want to describe the behaviour of the output, g on an arbitrary name-intension. Therefore let σ be a name intension (a function from indices to objects), and consider $g(\sigma)$, which must be a sentence intension (a function from indices to truth-values). Stipulate that $g(\sigma)$ will take the value T at an index i iff $f(i') = T$, where i' is an index is like i except that the object it assigns the free variable matches the object delivered by σ. (I.e. $f(i') = T$) where i' is that index differs from i if at all only at the nth entry in the variable-assignment co-ordinate of i; where it there coincides with $\sigma(i)$.)

[2] I.e. if there are denumerably many variables of category N, $v_1,\ldots,v_n\ldots$ then a variable assignment is a mapping from the natural numbers to objects.

variable assignment maps variables to intensions, anything in the transitive closure of the intension is in the transitive closure of the variable assignment. In particular, a given variable assignment is in its own transitive closure—a violation of the foundation axiom (cf. Lewis, 1970a, postscript).[3]

The plan for the remainder of this note is the following. First, we generalize general semantics by making a distinction between *compositional* and *Carnapian* indices, and relate this to the double-indexed semantics and the Carnapian type theory described in §§5.1 and 5.4 respectively. We note that treating variable assignments as Carnapian indices resolves the worries about Foundation, but does not allow an easy generalization of Lewis' treatment of binders. Setting the question of how variables are to be handled aside for the moment, we note that we already have the resources for proving sentential invariance under permuted reference-schemes in the general setting, and that Carnapian indices always allow the analogue of radical indexical inscrutability. This runs against recent arguments by McGee (2005a), and so we take time to explain how our setting—and in particular, the treatment of variable assignments as indices—undermines McGee's case. Finally, we turn back to the question of how λ-terms are to be handled. I show how to extend the permutation argument to (a version of) Cresswell's λ-categorial framework.

Generalizing general semantics

Let us initially divide indices within an intensional type theory into two sorts: Carnapian indices and compositional indices. Now define two sorts of intension for an expression. Appropriate α-intensions will be characterized recursively as compositional intensions are in Lewis' treatment, and as appropriate extensions are in an extensional type theory. Having chosen α-intensions for basic categories, we define appropriate α-intensions for derived categories in a way isomorphic to the build up of the categories themselves. In particular, an expression of category S/N will be assigned a function from name-α-intensions (functions from compositional indices to objects) to sentence-α-intensions (functions from compositional indices to truth-vales).

Next, define appropriate β-intensions (which will play the role of semantic values) over the α-intensions just constructed, paralleling the moves made in the Carnapian case. A β-intension for a derived category will then be a function from Carnapian indices to appropriate extensions. A name will therefore be assigned as semantic value a function from Carnapian indices to name-α-intensions; i.e. a function from Carnapian indices to (a function from compositional indices to truth-values). S/N expressions will be assigned as semantic value a function from Carnapian indices to (a mapping from name-α-intension to sentence α-intensions).

Various of the frameworks discussed in Chapter 5 can now be categorized. The Carnapian intensional type theory (§5.4) is a case where all indices—variable assignments, possible worlds, contexts etc—are all treated as compositional indices. The α intensions are just extensions, and the β intensions are 'Carnapian intensions'. The single-indexed Lewisian framework of compositional intensional type theory is a case where all these indices are compositional indices. The α intensions are Lewis' compositional intensions, and since there are no Carnapian indices we can identify β-intensions with the α-intensions.

The double-indexed general semantics that I put forward for the purposes of the argument for radical indexical inscrutability is of the mixed kind. All the indices bar that of context are compositional, but the context index is Carnapian. In this framework, the α intensions are Kaplanian 'contents' (handled compositionally) whereas the β intensions are Kaplanian 'characters': functions from contexts to contents.

How should we decide whether to treat an index as compositional or as Carnapian? Lewis' argument in favour of compositional indices rests on what he calls 'intensional predicates' such as (allegedly) "is

[3] Montague's framework in (1970), by contrast, is based on Carnapian intensions (albeit with departures), and so Lewis' problems with variable-assignments for derived types do not afflict him.

rising" where whether the concatenation of the predicate with a name is true at a world depends on the intension of a name, not just its extension. (Other phenomena calling for such treatment include adjectives such as 'good', 'fake' etc.)

It seems we should treat possible worlds as a compositional index, if only as a matter of prudence. For example, there seems no parallel argument against treating context in a Carnapian way.

Suppose we treat variable assignments as Carnapian. Immediately, some of the problems for Lewis' picture are addressed. For we can now take variable assignments to map variables to appropriate α intensions; but α-intensions do *not* include variable assignments in their transitive closure, so there is no threat of violating Foundation. This is definite progress: we do indeed have the basic framework for handling variables of all categories.

There is a bug. Lewis' definition of binders *treats them as functions of the* β*-intensions* of the sentences they attach to (given that variable assignments are now Carnapian). For what λx does when attached to a sentence θ is to create a function f, such that the value of $f(\alpha)$ at a Carnapian index is defined in terms of $|\theta|$ at some *other* Carnapian index (i.e. some other variable assignment). Although we have no problems with Foundation, *prima facie* we still cannot define binders, if we are to limit ourselves to the Lewisian functional-application treatment of the semantic significance of concatenation.[4] I will shortly follow Lewis in moving to Cresswell's (1973) λ-categorial framework, and show that permutation arguments still go through in that setting. Significantly for what follows, we can still treat variable assignments as Carnapian indices in the way just sketched.

I will now discuss how this treatment of variables undermines some objections to permutation arguments due to Vann McGee.

The inscrutability results and variables

We should note that we already have the materials to prove radical interpretation results for the 'mixed' setting. The theorem of §5.1.2 showed that α-intensions were invariant under permutations, and in extending this result to a double-indexed semantics, we implicitly showed that this holds also for β-intensions. Further, the cut-and-shunt argument of §5.1 for radical indexical inscrutability relies only on the Carnapian character of the context index. For any Carnapian index i, we can establish radical i-inscrutability by exactly the same procedure there used. In particular, this holds for the limiting case, where all indices are Carnapian, as was noted in passing in §5.4.

McGee (2005a) holds that variables are the Achilles heel of one version of permutation arguments for radical inscrutability. The argument he considers is the one where different permutations are applied to each world—effectively, this is the 'cut-and-shunt' technique, this time for radical *intensional* inscrutability we discussed in §5.4.[5] McGee's first complaint is that if different permutations are chosen in different worlds, then names will no longer be *rigid designators*. I am confused as to how this observation could *undermine* permutation arguments. To note that radical intentional inscrutability will undermine the rigidity of proper names is no more an objection to inscrutability arguments than it is to note that straightforward radical inscrutability can associate names with objects with which tokenings of the name bear no causal connection. What would be needed is some *independent* constraint on acceptable interpretation: respectively, that proper names must be causally connected to their referents, or that they must be rigid.[6]

[4] Achille Varzi, in personal communication, maintains that binders can be handled within a pure categorial language. The outline of his treatment is in Varzi (2002). If Varzi is right, then the following extension to λ-categorial language is unnecessary. I hope in future work to be in a position to evaluate Varzi's contention, the details of which are given in Varzi (1999).

[5] As there mentioned, this kind of 'cut-and-shunt' move is first made in the appendix to Putnam (1980).

[6] Even if we made the case that everything of type N has to be rigid (as for example, Evans (1982) tries to do, based on a principle of simplicity in semantic theorizing), another premiss that would be required is that the elements of English we ordinarily take to be names are indeed proper names. Ramsey famously worries about this, holding that names might be

A stronger argument that McGee offers is the following. Suppose that humans are essentially animals. Then the following should be false: there is something which is a human, and is such that it is possibly a chair. Given radical intensional inscrutability, we can let the permutation chosen for the actual world be the identity mapping, and the permutation for some non-actual world w be one that maps Tony Blair to his chair and vice versa. Given the way that the deviant interpretation is constructed in w, Tony Blair will fall under 'chair' at w. On a standard treatment of variable assignments, an assignment that maps x to Tony Blair will render true the open sentence x is human and possibly, x is a chair. As McGee puts it, variables are *automatically* rigid, so the permutation argument (in the current formulation) gives the wrong results when we consider quantified modal formulae.

This generates a puzzle: I have claimed to prove that categorial languages, of a sort that is able to handle all the quantification one could wish for, is sententially invariant under the kind of permutations that McGee is considering.[7] Yet McGee has produced an apparent counterexample. The tension is defused, however, when we realize that our setting does *not* treat variables assignments in the standard way. From the perspective of our system, variables are simply a term of category N, not distinct, from a formal point of view, from an ordinary proper name. Under the intended interpretation, the extension of this term depends solely on the variable assignment, and is indifferent to what world-index is selected. *This is not the case on the permuted interpretation.* Following through the way this was to be constructed, we see that *any* expression in category N at world w will refer to the ϕ-image of its intended referent. Thus variables too will become non-rigid under radical intensional inscrutability.[8] Hence our setting is safe from McGee's alleged counterexample.[9]

The permutation argument in a λ-categorial setting.

While keeping the treatment of variable assignments as Carnapian indices, let us move (as Lewis (1970a) suggests) to a λ-categorial setting based on that described by Cresswell (1973). Here, λ expressions are syncategorematic, and we need to lay down specific semantic axioms governing them, as follows:

> **λ-axiom**
> Where x is in category σ and θ is of type t, $|\lambda x\theta|$ is the function f of type $\langle t,\sigma\rangle$, such that for any $a \in \sigma$, $f(a)(i) = |\theta|(i')$, where i' and i match everywhere except possibly on the value assigned to x, where i' assigns a. (We write this $i' = i(x/a)$)

The question now is whether our permutation results can be extended to this revised setting.

Now the proof as we have given in §5.1 still stands: it assigned a value to each expression according to its category, and showed how these fit together to give an overall compositional interpretation. Now, however, there is the worry that this interpretation will not verify the λ-axiom we just laid down—that $|\lambda x\theta|_\phi$ and to $|\theta|$ will not be connected in the way just specified. In fact, this worry is well-placed right. In order to get a permuted interpretation to work, we have to include in our semantic theory a *different* axiom for λ, as follows:

> **ϕ-variant λ-axiom**
> Where x is in category σ and θ is of type t, $|\lambda x\theta|_\phi$ is the function f of type $\langle t,\sigma\rangle$, such that for

re-construed as second-level predicates (i.e. of category (S/(S/N)). Indeed, Montague handled proper names in exactly this way.

[7] Lewis' treatment of objectual quantification in his pure categorial language would suffice to make this point, given our results in Chapter 5, even if the extension to λ-categorial languages below was not successful.

[8] Note that the value of a variable x_n at variable-assignment v will no longer simply be the object to which the variable-assignment maps the appropriate n.

[9] Note also that McGee's claims are in tension with the proofs given in the appendix to Hale and Wright (1997b). The framework there used (deriving from Benson Mates) effectively treats variables as temporary names. Hence, when the permutation argument is proved, variables are subject to the same 'twisting' as are names, and so, in the case McGee considers, will no longer be rigid.

any $a^{\phi} \in \sigma$, $f(a^{\phi})(i) = |\theta|_{\phi}(i')$, where i' and i are Carnapian indices that match everywhere except possibly on the value assigned to x, where i' assigns a (i.e. $i' = i(x/a)$).

We can now proceed to show that nothing goes wrong with the permutation result. We need to show that our axiom is vindicated by the permuted-variant interpretation. The crucial clause for our purposes is the following:

> If f is an appropriate intension for $C/C_1 \ldots C_n$,
> then $f^{\phi} : r^{\phi} \mapsto s^{\phi}$ iff $f : r \mapsto s$.

Of course, we now modify this since we have Carnapian indices in place. What we need in general is the following recursive clause:

> If f is an appropriate β-intension for $C/C_1 \ldots C_n$, and i is a specification of the Carnapian indices, then $f^{\phi}(i) : r^{\phi}(i) \mapsto s^{\phi}(i)$ iff $f(i) : r(i) \mapsto s(i)$ where $r^{\phi} = g$ and $s^{\phi} = h$

In the case at hand, the functions in question will be mapping intensions onto sentence intensions. We already know that sentence intensions are invariant across the original and permuted interpretation. What we need to show, therefore, is that $|\lambda x\theta|(r)(i) = |\lambda x\theta|_{\phi}(r^{\phi})(i)$. But this follows immediately from the permuted λ-axiom. By the λ-axiom:

$$|\lambda x\theta|(r)(i) = |\theta|(i(x/r)). \tag{$\dagger$}$$

If we applied the old rule to the permuted interpretation, $|\lambda x\theta|_{\phi}(r^{\phi})(i) = |\theta|_{\phi}(i(x/r^{\phi}))$. But since we are using the new ϕ-variant rule for λ, we have

$$|\lambda x\theta|_{\phi}(r^{\phi})(i) = |\theta|_{\phi}(i(x/r)). \tag{$\ddagger$}$$

$|\theta|$ and $|\theta|_{\phi}$ are identical (being of category t), so ($\ddagger$) gives us $|\lambda x\theta|_{\phi}(r^{\phi})(i) = |\theta|(i(x/r))$. Since this is identical to $|\lambda x\theta|(r)(i)$ (by $\dagger$), we have our result.[10]

.

[10] Strictly, we would induct on the number of occurrences of λ in θ.
To prove the result for the fully general λ-categorial language, where θ can be of any category, further work would be needed. I assume that this will pose no new problems, and hope to give the fully general result in future work.

Bibliography

Ajdukiewicz, K. (1935). ‘Die syntaktische konnexität’. *Studia Philosophica*, **1**, 1–27. English translation: ‘Syntactic Connection’, in McCall S.(ed.), *Polish Logic 1920-1939* (Oxford University Press: Oxford, 1967), 207-231.

Armstrong, D. (1980). ‘Identity through time’. In P. van Inwagen, editor, *Time and Cause*. D. Reidel, Dordrecht.

Armstrong, D. (1986). ‘In defense of structural universals’. *Australasian Journal of Philosophy*, **64**, 85–88.

Armstrong, D. M. (1978a). *Nominalism and Realism: Universals and scientific realism vol I*. Cambridge University Press, Cambridge.

Armstrong, D. M. (1978b). *A Theory of Universals: Universals and scientific realism vol II*. Cambridge University Press, Cambridge.

Armstrong, D. M. (1983). *What is a Law of Nature?* Columbia University Press, New York.

Armstrong, D. M. (1989). *Universals: An opinionated introduction*. Westview Press, Boulder: Colorado.

Avramides, A. (1997). ‘Intention and convention’. In C. Wright and B. Hale, editors, *A Companion to the Philosophy of Language*, pages 60–86. Blackwell, Oxford.

Ayers, M. (1974). ‘Individuals without sortals’. *Canadian Journal of Philosophy*, **4**(1), 113–148.

Barnes, E. (2005). ‘Vagueness in sparseness: a study in property ontology’. *Analysis*, **65**.

Beall, J. C. and van Fraassen, B. (2003). *Possibilities and Paradox*. Oxford University Press, Oxford.

Benacerraf, P. (1965). ‘What numbers could not be’. In P. Benacerraf and H. Putnam, editors, *Philosophy of Mathematics: Selected readings*, pages 272–294. Cambridge University Press, Cambridge (1983), second edition. First published in *Philosophical Review* 74 (1965): 47–73.

Benacerraf, P. and Putnam, H., editors (1983). *Philosophy of Mathematics: Selected readings*. Cambridge University Press, Cambridge, second edition.

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**, 430–448. Reprinted in Boolos *Logic, Logic and Logic* (Harvard University Press, Cambridge: 1998) pp.54-72.

Boolos, G. (1998). *Logic, logic, and logic*. Harvard University Press, Cambridge, MA. Edited by Richard Jeffrey.

Braddon-Mitchell, D. (2001). ‘Lossy laws’. *Noûs*, **35**(2), 260–277.

Braddon-Mitchell, D. and Jackson, F. (1996). *Philosophy of Mind and Cognition*. Blackwell, Oxford.

Brandom, R. (1996). 'The significance of complex numbers for Frege's philosophy of mathematics'. *Proceeding of the Aristotelian Society*, pages 293–315.

Brock, S. (1993). 'Modal fictionalism: A response to Rosen'. *Mind*, **102**(405), 147–150.

Burgess, J. and Rosen, G. (1997). *A Subject with no Object*. Oxford University Press, Oxford.

Burns, L. C. (1991). *Vagueness: an investigation into natural language and the sorites paradox*. Oxford University Press, New York.

Campbell, J. (1994). *Past, Space and Self*. MIT Press, Cambridge, MA.

Campbell, J. (1997). 'Precis of *Past, Space and Self* and replies to critics'. *Philosophy and Phenomenological Research*, **57**(3), 633–634, 655–670.

Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press, Oxford.

Cohen, P. J. (1963). 'The independence of the continuum hypothesis: I'. *Proceedings of the National Academy of Sciences U.S.A.*, **50**, 1143–1148.

Cohen, P. J. (1964). 'The independence of the continuum hypothesis: II'. *Proceedings of the National Academy of Sciences U.S.A.*, **51**, 105–110.

Cohen, P. J. (1966). *Set theory and the continuum hypothesis*. Benjamin, New York.

Colyvan, M. (2003). *The Indispensibility of Mathematics*. Oxford University Press, Oxford.

Cresswell, M. (1973). *Logics and Languages*. Methuen, London.

Cresswell, M. (2004). 'Adequacy conditions for counterpart theory'. *Australasian Journal of Philosophy*, **82**, 28–41.

Davidson, D. (1965). 'Theories of meaning and learnable languages'. In Y. Bar-Hillel, editor, *Proceedings of the 1964 Inrternational Congress for Logic, Methodology and Philosophy of Science*. North Holland, Amsterdam. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.3–16.

Davidson, D. (1967). 'Truth and meaning'. *Synthese*, **17**, 304–23. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.17–36.

Davidson, D. (1973). 'Radical interpretation'. *Dialectica*, **27**, 313–28. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.125–140.

Davidson, D. (1974). 'Belief and the basis of meaning'. *Synthese*, **27**, 309–23. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.141–154.

Davidson, D. (1977). 'Reality without reference'. *Dialectica*, **31**, 247–53. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.215–226.

Davidson, D. (1979). 'The inscrutability of reference'. *The Southwestern Journal of Philosophy*, pages 7–19. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.227–242.

Davidson, D. (1980). 'Towards a unified theory of thought, meaning and action'. *Grazer Philosophische Studien*, **11**, 1–12. Reprinted in Davidson, *Problems of Rationality* (Oxford University Press, Oxford: 2004) pp.151-166.

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

Davidson, D. (1997). 'Indeterminism and anti-realism'. In C. B. Kulp, editor, *Realism/Antirealism and epistemology*. Rowman and Littlefield, Lanham. Reprinted in Davidson, *Problems of Rationality* (Oxford University Press, Oxford: 2004) pp. 69-84.

Davidson, D. (2004). *Problems of Rationality*. Oxford University Press, Oxford.

Davies, M. (1981). *Meaning, Quantification, Necessity*. Routledge and Kegan Paul, London.

Davis, S. and Gillon, B. S. (2004). *Semantics: A reader*. Oxford University Press, Oxford.

Dorr, C. (2002). 'The simplicity of everything'. PhD Dissertation, Princeton University. Online at `http://www.pitt.edu/~csd6/`.

Dorr, C. (2003). 'Vagueness and ignorance'. *Philosophical Perspectives*, **17**, 83–114.

Dorr, C. (2004). 'Non-symmetric relations'. In D. Zimmerman, editor, *Oxford Studies in Metaphysics*. Oxford University Press, Oxford.

Dorr, C. (forthcoming 2005). 'There are no abstract objects'. In J. Hawthorne, T. Sider, and D. Zimmerman, editors, *Blackwell Great Debates in Metaphysics*. Blackwell, Oxford.

Dorr, C. and Rosen, G. (2002). 'Composition as a fiction'. In *The Blackwell Guide to Metaphysics*. Blackwell, Oxford.

Dowty, D. R. (1979). *Word Meaning and Montague Grammar*. D Reidel, Dordrecht, Holland.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. The David Hume Series. CSLI Publications, Stanford.

Dummett, M. A. E. (1973). *Frege: Philosophy of Language*. Duckworth, London.

Dummett, M. A. E. (1975). 'Wang's paradox'. *Synthese*, **30**, 301–324. Reprinted in Keefe and Smith (eds) *Vagueness: A Reader* (MIT Press: 1997) pp.99–119.

Dummett, M. A. E. (1976). 'What is a theory of meaning? Part II'. In G. Evans and J. McDowell, editors, *Truth and Meaning: Essays in semantics*, pages 67–137. Oxford University Press, Oxford.

Dummett, M. A. E. (1991). *The Logical Basis of Metaphysics*. The William James lectures; 1976. Harvard University Press, Cambridge, Mass.

Edgington, D. (1995). 'On conditionals'. *Mind*, **104**, pp. 235–329.

Edgington, D. (1997). 'Vagueness by degrees'. In R. Keefe and P. Smith, editors, *Vagueness: A reader*. MIT Press, Cambridge, MA.

Eklund, M. (Draft, 2005). 'Vagueness and second level indeterminacy'. `http://spot.colorado.edu/~eklundm/levels.pdf`.

Elga, A. (2004). 'Infinitesimal chances and laws of nature'. *Australasian Journal of Philosophy*, **82**, 67–76.

Evans, G. (1975). ‘Identity and predication’. *Journal of Philosophy*, **LXXII**(13), 343–362. Reprinted in McDowell, (ed) *Gareth Evans: Collected Papers* (Clarendon Press, Oxford).

Evans, G. (1981). ‘Semantic theory and tacit knowledge’. In Holtzman and Leich, editors, *Wittgenstein: To follow a rule*. Routledge and Kegan Paul, London. Reprinted in McDowell, (ed) *Gareth Evans: Collected Papers* (Clarendon Press, Oxford) pp.322-42.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press, Oxford. Edited by John McDowell.

Ewald, W. B. (1996). *From Kant to Hilbert: A source book in the foundations of mathematics*. Oxford science publications. Clarendon, Oxford. Translated into English from multiple languages.

Field, H. H. (1972). ‘Tarski’s theory of truth’. *Journal of Philosophy*, **69**, 347–375. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 3-29.

Field, H. H. (1973). ‘Theory change and the indeterminacy of reference’. *Journal of Philosophy*, **70**, 462–81. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 177-198.

Field, H. H. (1974). ‘Quine and the correspondence theory’. *Philosophical Review*, **83**, 200–228. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 199-221.

Field, H. H. (1975). ‘Conventionalism and instrumentalism in semantics’. *Noûs*, **9**, 375–405.

Field, H. H. (1977). ‘Logic, meaning and conceptual role’. *Journal of Philosophy*, **74**(7), 379–409.

Field, H. H. (1978). ‘Mental representation’. *Erkenntnis*, **13**, 9–61. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 30-67.

Field, H. H. (1980). *Science without Numbers: A defence of nominalism*. Library of philosophy and logic. Blackwell, Oxford.

Field, H. H. (1989). *Realism, Mathematics and Modality*. Basil Blackwell, Oxford.

Field, H. H. (1994). ‘Deflationist views of meaning and content’. *Mind*, **103**, 249–85. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 332-360.

Field, H. H. (1998). ‘Some thoughts on radical indeterminacy’. *Monist*, **81**(2), 253–73. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 259-278.

Field, H. H. (2001a). ‘Attributions of meaning and content’. In *Truth and the Absence of Fact*. Oxford University Press, Oxford. First published in this volume.

Field, H. H. (2001b). *Truth and the Absence of Fact*. Oxford University Press, Oxford.

Field, H. H. (2005). ‘*Truth and the Absence of Fact*: Precis and replies to commentators’. *Philosophical Studies*, **124**, 41–44; 105–128.

Fine, K. (1975). ‘Vagueness, truth and logic’. *Synthese*, **30**, 265–300. Reprinted with corrections in Keefe and Smith (eds) *Vagueness: A reader* (MIT Press, Cambridge MA: 1997) pp.119-150.

Fine, K. (2003). ‘The problem of possibilia’. In M. Loux and D. Zimmerman, editors, *The Oxford Handbook of Metaphysics*, pages 161–179. Oxford University Press, Oxford.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Bradford, Cambridge, MA.

Fodor, J. (1993). *The Elm and the Expert: Mentalese and its semantics*. Bradford, Cambridge, MA.

Fodor, J. and Lepore, E. (1992). *Holism: A shoppers guide*. Blackwell, Oxford.

Forrest, P. (1986). 'Neither magic nor mereology: A reply to Lewis'. *Australasian Journal of Philosophy*, **64**, 89–91.

Foster, J. A. (1976). 'Meaning and truth theory'. In G. Evans and J. McDowell, editors, *Truth and Meaning: Essays in semantics*, pages 1–32. Clarendon Press, Oxford.

Frege, G. (1892). 'Über Sinn und Bedeutung'. *Zeitschrift für Philosophie und philosophische Kritik*, **100**, 25–50. Translation by Max Black in Geach and Black (1970): 56–78.

Garcia-Carpintero, M. (2000). 'A presuppositional account of reference-fixing'. *Journal of Philosophy*, **109-147**, 71–104.

Gibbard, A. (1975). 'Contingent identity'. *The Journal of Philosophical Logic*, **4**, 187–221.

Gödel, K. (1938). 'The consistency of the axiom of choice and of the generalized continuum-hypothesis'. *Proceedings of the National Academy of Sciences U.S.A.*, **24**, 556–557.

Gödel, K. (1940). *The Consistency of the Continuum-Hypothesis*. Princeton University Press, Princeton, NJ.

Greenough, P. (2003). 'Vagueness: A minimal theory'. *Mind*, **112**(446).

Grice, H. P. (1975). 'Logic and conversation'. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. New York Academic Press, New York. Reprinted in A. P. Martinich (ed) *The Philosophy of Language* fourth edition (Oxford University Press, Oxford: 2000) pp.165–175.

Gupta, A. and Martínez-Fernández, J. (2005). 'Field on the concept of truth: Comment'. *Philosophical Studies*, **124**, 45–58.

Hale, B. and Wright, C., editors (1997a). *A Companion to the Philosophy of Language*. Blackwell, Oxford.

Hale, B. and Wright, C. (1997b). 'Putnam's model-theoretic argument against metaphysical realism'. In C. Wright and B. Hale, editors, *A Companion to the Philosophy of Language*, pages 427–457. Blackwell, Oxford.

Hart, W. D. (1996). *The Philosophy of Mathematics*. Oxford readings in philosophy. Oxford University Press, Oxford.

Haslanger, S. (2003). 'Persistence through time'. In M. Loux and D. Zimmerman, editors, *The Oxford Handbook of Metaphysics*, pages 315–356. Oxford University Press, Oxford.

Hawley, K. (2001). *How Things Persist*. Oxford University Press, Oxford.

Heal, J. (1997). 'Radical interpretation'. In C. Wright and B. Hale, editors, *A Companion to the Philosophy of Language*. Blackwell, Oxford.

Heck, R. (forthcoming 2005a). 'Reason and language'. In C. MacDonald, editor, *McDowell and his Critics*.

Heck, R. (forthcoming 2005b). 'Use and meaning'. In *The Philosophy of Michael Dummett*.

Hellman, G. (1989). *Mathematics Without Numbers: Towards a modal-structural interpretation*. Clarendon Press, Oxford.

Henkin, L. (1949). 'The completeness of the first order logical calculus'. *The Journal of Symbolic Logic*, **14**, 159–166.

Henkin, L. (1950). 'Completeness in the theory of types'. *The Journal of Symbolic Logic*, **15**, 81–91.

Hirsch, E. (1993). *Dividing Reality*. Oxford University Press, New York.

Hodes, H. (1984). 'Logicism and the ontological commitments of arithmetic'. *Journal of Philosophy*, **81**, 123–149.

Hornsby, J. (2005). 'Semantic knowledge and practical knowledge: I'. *Proceedings of the Aristotelian Society: Supplementary Volume*, **supp. LXVIII**, 107–130.

Horwich, P. (1990). *Truth*. Basil Blackwell, Oxford.

Hyde, D. (1997). 'From heaps and gaps to heaps of gluts'. *Mind*, **106**, 641–60.

Jackson, F. (1997). 'Reference and descriptivism defended'. Draft downloaded from `http://philrsss.anu.edu.au/people-defaults/fcj/index.php3` on 13/05/03.

Jackson, F. (1998). *From Metaphysics to Ethics: A defence of conceptual analysis*. Oxford University Press, Oxford.

Jeffrey, R. (1964). 'Review of *Logic, Methodology and the Philosophy of Science*, ed. E. Nagel, P. Suppes and A. Tarski'. *Journal of Philosophy*, **61**, 79–88.

Jeffrey, R. (1965). *The Logic of Decision*. University of Chicago Press, Chicago and London, 2nd edition. Second edition published 1983.

Kamp, J. A. W. (1975). 'Two theories about adjectives'. In E. Keenan, editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press, Cambridge. Reprinted in Davis and Gillon (eds) *Semantics: A reader* (Oxford University Press, Oxford, 2004) pp.541-562.

Kaplan, D. (1989a). 'Afterthoughts'. In J. Almog, J. Perry, and H. Wettstein, editors, *Themes from Kaplan*, chapter 18, pages 564–614. Oxford University Press, New York; Oxford.

Kaplan, D. (1989b). 'Demonstratives'. In J. Almog, J. Perry, and H. Wettstein, editors, *Themes from Kaplan*, chapter 17, pages 481–563. Oxford University Press, New York; Oxford.

Kaplan, D. (1990). 'Words'. *Proceedings of the Aristotelian Society*, **supp LXIV**.

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

Keefe, R. and Smith, P. (1997). *Vagueness: A reader*. MIT Press, Cambridge, MA.

Kim, J. (1974). 'Noncausal connections'. *Noûs*, **8**, 41–52.

Kölbel, M. (2001). 'Two dogmas of Davidsonian semantics'. *The Journal of Philosophy*, **XCVIII**(12).

Kreisel, G. (1967). 'Informal rigour and completeness proofs'. In I. Lakatos, editor, *Problems in the Philosophy of Mathematics*, pages 138–171. North Holland, Amsterdam.

Kripke, S. A. (1980). *Naming and Necessity*. Blackwell, Oxford, revised and enlarged edition. Originally published in Donald Davidson and Gilbert Harman (eds) *Semantics of Natural Language*; Dordrecht: Reidel, 1972.

Kuratowski, K. (1921). 'Sur la notion de l'ordre dans la théorie des ensembles'. *Fundamenta Mathematicae*, **2**, 161–171.

Larson, R. K. and Ludlow, P. (1993). 'Interpreted logical forms'. *Synthese*, **95**, 305–356.

Larson, R. K. and Segal, G. (1995). *Knowledge of Meaning*. MIT Press, Cambridge, MA.

Lepore, E. and Ludwig, K. (2005). *Donald Davidson: Meaning, truth, language and reality*. Oxford University Press, Oxford.

Lewis, D. K. (1968). 'Counterpart theory and quantified modal logic'. *Journal of Philosophy*, **65**, 113–26. Reprinted with postscript in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 26–39.

Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press, Cambridge, MA.

Lewis, D. K. (1970a). 'General semantics'. *Synthese*, **22**, 18–67. Reprinted with postscript in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 189–229.

Lewis, D. K. (1970b). 'How to define theoretical terms'. *Journal of Philosophy*, **67**, 427–446. Reprinted in Lewis, *Philosophical Papers vol. I* (Oxford University Press, 1983) 78–95.

Lewis, D. K. (1973a). 'Causation'. *Journal of Philosophy*, **70**, 556–67. Reprinted with postscripts in Lewis, *Philosophical Papers vol. II* (Oxford University Press, 1986) 159-71.

Lewis, D. K. (1973b). *Counterfactuals*. Blackwell, Oxford.

Lewis, D. K. (1974a). 'Radical interpretation'. *Synthese*, **23**, 331–44. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 108–18.

Lewis, D. K. (1974b). ''tensions'. In M. K. Munitz and P. K. Unger, editors, *Semantics and Philosophy*, pages 241–70. New York University Press, New York. Reprinted with postscripts in Lewis, *Philosophical Papers vol. I* (Oxford University Press, 1983).

Lewis, D. K. (1975). 'Language and languages'. In *Minnesota Studies in the Philosophy of Science*, volume VII, pages 3–35. University of Minnesota Press. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 163-88.

Lewis, D. K. (1976). 'Survival and identity'. In A. O. Rorty, editor, *The Identities of Persons*. University of California Press. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 55–72.

Lewis, D. K. (1979a). 'Attitudes *de dicto* and *de se*'. *The Philosophical Review*, **88**, 513–43. Reprinted with postscript in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 133-55.

Lewis, D. K. (1979b). 'Counterfactual dependence and time's arrow'. *Noûs*, **13**, 455–76. Reprinted with postscript in Lewis, *Philosophical Papers II* (Oxford University Press, 1986) 32–51.

Lewis, D. K. (1979c). 'Scorekeeping in a language game'. *Journal of Philosophical Logic*, **8**, 339–59. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 233-49.

Lewis, D. K. (1980). 'Index, context and content'. In S. Kanger and S. Öhman, editors, *Philosophy and Grammar*, pages 79–100. Reidel, Dordrecht. Reprinted in Lewis, *Papers on Philosophical Logic* (Cambridge University Press, 1998) 21–44.

Lewis, D. K. (1983a). 'New work for a theory of universals'. *Australasian Journal of Philosophy*, **61**, 343–377. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 8–55.

Lewis, D. K. (1983b). *Philosophical Papers*, volume I. Oxford University Press, Oxford, New York.

Lewis, D. K. (1984). 'Putnam's paradox'. *Australasian Journal of Philosophy*, **62**(3), 221–36. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 56–77.

Lewis, D. K. (1986a). 'Against structural universals'. *Australasian Journal of Philosophy*, **64**, 25–46. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 78–107.

Lewis, D. K. (1986b). 'Causal explanation'. In *Philosophical Papers*, volume II, pages 214–240. Oxford University Press, Oxford, New York. First published in this collection.

Lewis, D. K. (1986c). *On the Plurality of Worlds*. Blackwell, Oxford.

Lewis, D. K. (1986d). *Philosophical Papers*, volume II. Oxford University Press, Oxford, New York.

Lewis, D. K. (1992). 'Meaning without use: Reply to Hawthorne'. *Australasian Journal of Philosophy*, **70**, 106–110. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 1999) 145–151.

Lewis, D. K. (1993). 'Many, but almost one'. In K. Campbell, J. Bacon, and L. Reinhardt, editors, *Ontology, Causality and Mind: Essays on the philosophy of D. M. Armstrong*. Cambridge University Press, Cambridge. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 164-82.

Lewis, D. K. (1994a). 'Humean supervenience debugged'. *Mind*, **103**, 473–90. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 224-47.

Lewis, D. K. (1994b). 'Reduction of mind'. In S. Guttenplan, editor, *A Companion to the Philosophy of Mind*, pages 412–31. Blackwell, Oxford. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 291–324.

Lewis, D. K. (1998). *Papers in Philosophical Logic*. Cambridge University Press, Cambridge.

Lewis, D. K. (1999). *Papers on Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

Lewis, D. K. (2000). *Papers on Ethics and Social Philosophy*. Cambridge University Press, Cambridge.

Lewis, D. K. (2003). 'Things qua truthmakers'. In H. Lillehammer and G. Rodriguez-Pereyra, editors, *Real Metaphysics*, pages 25–38. Routledge, London.

Lewis, D. K. (2004). 'Causation as influence'. In J. Collins, N. Hall, and L. Paul, editors, *Causation and counterfactuals*, pages 75–106. MIT Press, Cambridge, MA. Extended version of "Causation as Influence", *Journal of Philosophy* 97 (2000), 182–97.

Lewis, D. K. (typescript). 'Ramseyan humility'.

Lipton, P. (1991). *Inference to the Best Explanation*. Routledge, London, 2nd edition.

Loewer, B. (2005). 'On Field's *Truth and the Absence of Fact*: Comment'. *Philosophical Studies*, **124**, 59–70.

MacFarlane, J. G. (2000). *What does it mean to say that logic is formal?* University of Pittsburgh, Pittsburgh. PhD Dissertation. Available online at: `http://philosophy.berkeley.edu/macfarlane/diss.html`.

Mackie, J. L. (1976). *Problems from Locke*. Clarendon Press, Oxford.

Martinich, A. P. (2000). *The Philosophy of Language*. Oxford University Press, Oxford, fourth edition.

McCawley, J. (1980). *Everything that Linguistics have always wanted to Know about Logic (but were ashamed to ask)*. University of Chicago Press, Chicago.

McDowell, J. (1978). 'Physicalism and primitive denotation: Field on Tarski'. *Erkenntnis*, **13**, 131–52. Reprinted in McDowell, *Meaning, Knowledge and Reality* (Harvard University Press, 1998) pp.132-156.

McDowell, J., editor (1985). *Gareth Evans: Collected papers*. Oxford University Press, Oxford.

McGee, V. (2005a). "Inscrutability and its discontents". *Noûs*, **39**, 397–425.

McGee, V. (2005b). 'Two conceptions of truth?: Comment'. *Philosophical Studies*, **124**, 71–104.

McGee, V. and McLaughlin, B. (1994). 'Distinctions without a difference'. *Southern Journal of Philosophy*, **supp XXXII**, 203–251.

Melia, J. (1995). 'On what there's not'. *Analysis*, **55**, 223–9.

Merchant, J. (2005). "Fragments and ellipsis". *Linguistics and Philosophy*, **27**, 661–738.

Merrill, G. H. (1980). 'The model-theoretic argument against realism'. *Philosophy of Science*, **47**, 69–81.

Miller, A. (1997). 'Tacit knowledge'. In B. Hale and C. Wright, editors, *A Companion to the Philosophy of Language*, pages 146–174. Blackwell, Oxford.

Montague, R. (1970). 'Universal grammar'. In R. Thomason, editor, *Formal Philosophy: Selected papers of Richard Montague*. Yale University Press, New Haven and London.

Nolan, D. (1997a). 'Three problems for "strong" modal fictionalism'. *Philosophical Studies*, **87**, 259–275.

Nolan, D. (1997b). *Topics in the Philosophy of Possible Worlds*. Routledge, London.

Nolan, D. (2005). *David Lewis*. Acuman, London.

Nolan, D. (Summer 2002). 'Modal fictionalism'. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/sum2002/entries/fictionalism-modal/`.

Nolan, D. and O'Leary-Hawthorne, J. (1996). 'Reflexive fictionalisms'. *Analysis*, **56.1**, 23–32.

Parsons, C. (1974). 'Sets and classes'. *Noûs*, **8**, 1–12.

Parsons, C. (1990). 'The structuralist view of mathematical objects'. *Synthese*, **84**(3), 303–46. Reprinted in Hart (ed) *The Philosophy of Mathematics* (Oxford University Press, Oxford, 1996) pp.97–130.

Parsons, J. (2000). 'Must a four-dimensionalist believe in temporal parts?'. *Monist*, **83**(3), 399–418.

Parsons, J. (2005a). 'I am not now, nor have I ever been, a turnip'. *Australasian Journal of Philosophy*, **83**, 289–306.

Parsons, J. (Draft 2005b). 'Theories of location'. Available online at `http://weka.ucdavis.edu/~jp30/papers/`.

Partee, B. H. (1996). 'The development of formal semantics in linguistic theory'. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory*, pages 11–38. Blackwell, Oxford.

Perry, J. (1977). 'Frege on demonstratives'. *Philosophical Review*, **86**, 474–97. Reprinted in Yourgrau (ed) *Demonstratives* (Oxford University Press, Oxford: 1990) pp.50-70.

Pettit, D. (2002). 'Why knowledge is unnecessary for understanding language'. *Mind*, **111**, 519–550.

Prosser, S. (forthcoming 2005). 'Cognitive dynamics and indexicals'. *Mind and Language*.

Putnam, H. (1978a). 'Realism and reason'. In *Meaning and the moral Sciences*, pages 123–140. Routledge and Kegan Paul, London.

Putnam, H. (1978b). 'Reference and understanding'. In *Meaning and the moral Sciences*, pages 97–123. Routledge and Kegan Paul, London.

Putnam, H. (1980). 'Models and reality'. *The Journal of Symbolic Logic*, **45**(3), 421–444. Reprinted in Benacerraf and Putnam (eds.) *Philosophy of Mathematics: Selected readings*, second edition (Cambridge University Press, Cambridge: 1983).

Putnam, H. (1981). *Reason, Truth and History*. Cambridge University Press, Cambridge.

Quine, W. V. (1953). *From a Logical Point of View*. Oxford University Press, Oxford.

Quine, W. V. (1960). *Word and Object*. MIT Press, Cambridge, Mass.

Quine, W. V. (1964). 'Ontological reduction and the world of numbers'. *Journal of Philosophy*, **61**. Reprinted with substantial changes in Quine, *The Ways of Paradox and Other Essays: Revised and enlarged edition* (Harvard University Press, Cambridge, MA and London, 1976) pp.212—220.

Quine, W. V. (1976a). *The Ways of Paradox and Other Essays: Revised and enlarged edition*. Harvard University Press, Cambridge, MA and London, second edition. First edition published 1966.

Quine, W. V. (1976b). 'Worlds away'. *Journal of Philosophy*, **73**. Reprinted in Quine, *Theories and Things* (Harvard University Press, Cambridge MA and London, 1981) pp.124-128.

Quine, W. V. (1981). *Theories and Things*. Harvard University Press, Cambridge, Mass; London.

Rayo, A. (Draft, 2004). 'A metasemantic account of vagueness'. Available online at `http://philosophy2.ucsd.edu/~arayo/`.

Rayo, A. and Uzquiano, G. (1999). 'Towards a theory of second-order consequence'. *The Notre Dame Journal of Formal Logic*, **40**(3), 315–325.

Rayo, A. and Yablo, S. (2001). 'Nominalism through de-nominalization'. *Noûs*, **35**(1), 74–92.

Richard, M. (1997). 'Inscrutability'. *Canadian Journal of Philosophy*, **supp vol. 23**, 197–211.

Rosen, G. (1990). 'Modal fictionalism'. *Mind*, **99**, 327–354.

Rosen, G. (1993). 'A problem for fictionalism about possible worlds'. *Analysis*, **53**, 71–81.

Rothschild, D. and Leuenberger, S. (Draft 2005). 'Reference magnetism: from foundational semantics to structuralism'. As presented to BW4: The Origins of Reference. Fourth Barcelona Workshop on Topics in the Theory of Reference. May 2005.

Rumfitt, I. (1995). 'Truth conditions and communication'. *Mind*, **104**, 827–862.

Salmon, N. (1998). 'Nonexistence'. *Noûs*, **32**, 277–319.

Schaffer, J. (2003). 'Is there a fundamental level?'. *Noûs*, **37**, 498–517.

Schaffer, J. (2004). 'Two conceptions of sparse properties'. *Pacific Philosophical Quarterly*, **85**, 92–102.

Schaffer, J. (2005). 'Quiddistic knowledge'. *Philosophical Studies*, **123**, 1–32. Reprinted in Jackson and Priest (eds) *Lewisian Themes* (Oxford University Press: Oxford, 2005).

Schiffer, S. R. (1972). *Meaning*. Oxford University Press, Oxford.

Schwarz, W. (2005b). 'Parts and counterparts/'Semantic values and rabbit pictures'/'Why we need more intensions'. Internet Publication; Accessed 14/08/05. Archived discussion of Weblog. Respective permalinks: `http://www.umsu.de/wo/archive/2005/03/07/Parts_and_Counterparts`; `http://www.umsu.de/wo/archive/2005/03/31/Semantic_Values_and_Rabbit_Pictures`; `http://www.umsu.de/wo/archive/2005/04/26/Why_we_need_more_intensions`.

Schwarz, W. (Draft 2005a). 'Parts and counterparts'. `http://www.umsu.de/words/parts.pdf`.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, New York.

Shapiro, S. (1991). *Foundations without Foundationalism: A case for second-order logic*. Oxford logic guides ; 17. Clarendon Press, Oxford.

Sider, T. (1995). 'Sparseness, immanence, and naturalness'. *Noûs*, **29**, 360–377.

Sider, T. (1996a). 'All the world's a stage'. *Australian Journal of Philosophy*, **74**, 433–453.

Sider, T. (1996b). 'Naturalness and arbitrariness'. *Philosophical Studies*, **81**, 283–301.

Sider, T. (2001). *Four-dimensionalism*. Oxford University Press, Oxford.

Sider, T. (2002). 'The ersatz pluriverse'. *Journal of Philosophy*, **98**, 279–315.

Sider, T. (2003). 'Reductive theories of modality'. In M. Loux and D. Zimmerman, editors, *The Oxford Handbook of Metaphysics*, pages 180–210. Oxford University Press, Oxford.

Simons, P. (1987). *Parts: A study in ontology*. Oxford University Press, Oxford.

Soames, S. (1989). 'Semantics and semantic competence'. *Philosophical Perspectives*, **3**, 575–596.

Stainton, R. J. (1998). 'Quantifier phrases, meaningfulness "in isolation" and ellipsis'. *Linguistics and Philosophy*, **21**, 846–865.

Stalnaker, R. (1978). 'Assertion'. *Syntax and Semantics*, **9**, 315–332. Reprinted in Stalnaker, *Context and Content* (Oxford University Press, Oxford, 1999) pp.78-95.

Stalnaker, R. (1984). *Inquiry*. MIT Press, Cambridge, MA.

Stalnaker, R. (1987). 'Semantics for belief'. *Philosophical Topics*, **15**. Reprinted in Stalnaker, *Context and Content* (Oxford: Oxford University Press, 1999).

Stalnaker, R. (1988). 'Belief attribution and context'. In R. Grimm and D. Merrill, editors, *Contents of Thought*. University of Arizona Press, Tucson. Reprinted in Stalnaker, *Context and Content* (Oxford University Press, Oxford, 1999) pp.150–166.

Stalnaker, R. (1991). 'The problem of logical omniscience: I'. *Synthese*, **89**, 425–440. Reprinted in Stalnaker, *Context and Content* (Oxford: Oxford University Press, 1999).

Stalnaker, R. (1997). 'Reference and necessity'. In C. Wright and B. Hale, editors, *A Companion to the Philosophy of Language*, pages 534–554. Blackwell, Oxford.

Stalnaker, R. (1999a). *Context and Content*. Oxford University Press, Oxford.

Stalnaker, R. (1999b). 'The problem of logical omniscience: II'. In *Context and Content*, pages 255–74. Oxford University Press, Oxford. First published in this volume.

Stalnaker, R. (2001). 'On considering a possible world as actual'. *Proceedings of the Aristotelian Society Supplementary Volume*, **75**, 141–156.

Stanley, J. (2001). 'Hermenutic fictionalism'. In French and Wettstein, editors, *Midwest Studies XXV: Figurative Language*. Blackwell, Oxford.

Stanley, J. (2002). 'Modality and what is said'. *Philosophical Perspectives*, **16**, 321–344.

Sterelny, K. (1990). *The Representational Theory of Mind*. Blackwell, Oxford.

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. Methuen, London.

Thomason, R., editor (1974a). *Formal Philosophy: Selected papers of Richard Montague*. Yale University Press, New Haven and London.

Thomason, R. (1974b). 'Introduction'. In R. Thomason, editor, *Formal Philosophy: Selected papers of Richard Montague*. Yale University Press, New Haven, CT.

Unger, P. (1980). 'The problem of the many'. *Midwest Studies in Philosophy*, **5**, 411–67.

van Fraassen, B. (1989a). *Laws and Symmetry*. Clarendon Press, Oxford.

van Fraassen, B. (1989b). *The Scientific Image*. Clarendon Press, Oxford.

van Fraassen, B. (1997). 'Putnam's paradox: metaphysical realism revamped and evaded'. *Noûs*, **supp**, 17–42.

van Inwagen, P. (1990). *Material Beings*. Cornell University Press, Ithaca and London.

Varzi, A. (1999). *An Essay in Universal Semantics*. Kluwer, Dordrecht.

Varzi, A. (2002). 'On logical relativity'. *Philosophical Issues*, **12**, 197–219.

Wallace, J. (1977). 'Only in the context of a sentence do words have any meaning'. In P. French and T.E. Uehling, Jr., editors, *Midwest Studies in Philosophy 2: Studies in the Philosophy of Language*. University of Minnesota Press, Morris.

Weatherson, B. (2003). 'Many many problems'. *Philosophical Quarterly*, **53**, 481–501.

Weatherson, B. (2004). 'The problem of the many'. Fall 2004 edition edition. `http://plato.stanford.edu/archives/fall2004/entries/problem-of-many/`.

Weatherson, B. (Draft 2002). 'Vagueness and pragmatics'. Available at `http://brian.weatherson.net/Ch_8.pdf`.

Wedgewood, R. (2002). 'The aim of belief'. *Philosophical perspectives*, **16**, 267–297.

Wetzel, L. (1989). 'That numbers could be objects'. *Philosophical Studies*, **56**, 273–292.

Wiener, N. (1914). 'Simplification of the logic of relations'. *Proceedings of the Cambridge Philosophical Society*, **17**, 387–390. Reprinted in van Heijenoort (1967): pp.224–7.

Wiggins, D. (1980). *Sameness and Substance*. Blackwell, Oxford.

Williams, J. R. G. (forthcoming, 2006a). 'An argument for the many'. *Proceedings of the Aristotelian Society*.

Williams, J. R. G. (forthcoming 2006b). 'Eligibility and inscrutability'. *Philosophical Review*.

Williamson, T. (1994). *Vagueness*. Routledge, London.

Williamson, T. (1997). 'Sense, validity and context'. *Philosophy and Phenomenological Research*, **57**(3), 649–654.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press, Oxford.

Williamson, T. (2003). 'Everything'. *Philosophical Perspectives*, **17**, 415–465.

Wright, C. (1981). 'Rule-following, objectivity and the theory of meaning'. In C. Leich and S. Holtzman, editors, *Wittgenstein: To follow a rule*, pages 99–117. Routledge, London.

Wright, C. (1987). 'Theories of meaning and speaker's knowledge'. In *Realism, Meaning and Truth*, pages 204–38. Blackwell, Oxford.

Wright, C. (1997). 'The indeterminacy of translation'. In C. Wright and B. Hale, editors, *A Companion to the Philosophy of Language*, pages 397–426. Blackwell, Oxford.

Yablo, S. (2001). 'Go figure: A path through fictionalism'. *Midwest Studies in Philosophy*, **25**, 72–102.

Yourgrau, P. (1990). *Demonstratives*. Oxford University Press, Oxford.

Zermelo, E. (1930). 'Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre'. *Fundamenta Mathematicae*, **16**, 29–47. Translated by Michael Hallett in Ewald (1996), pp.1219–33.

Zilber, B. (2000). Lecture course on Model Theory, Oxford University. `http://www.maths.ox.ac.uk/~zilber/lect.pdf`.