Order Number 9119625

Extensionalist semantics after Quine

Edwards, Michael Charles, Ph.D.
City University of New York, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

EXTENSIONALIST SEMANTICS AFTER QUINE

by

MICHAEL CHARLES EDWARDS

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1991

This manuscript has been read and accepted by the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

December 19, 1990
Date

Arthur W Hall
Chair of Examining Committee

December 19, 1990
Date

Arthur W Hall
Executive Officer

Prof. Jerrold J. Katz

Prof. Arnold Koslow

Prof. Richard L. Mendelsohn
Supervisory Committee

The City University of New York

Abstract

EXTENSIONALIST SEMANTICS AFTER QUINE

by

Michael Charles Edwards

Advisor: Professor Jerrold J. Katz

Quine's work in semantics has sought to achieve a pair of results: by discrediting intensional entities and intensional logics, to show that extensionalist resources are the only ones available for the construction of a theory of meaning; and to establish that, given only the resources of extensionalism, semantic content is indeterminate beyond the yield of stimulus meaning. The present dissertation focuses upon the second of these issues.

After examining Quine's reasons for eschewing intensionalism and his case for indeterminacy, I proceed to assess the merits of five theories of meaning, those developed by Donald Davidson, David Lewis, Hilary Putnam, Gilbert Harman and Hartry Field. All these philosophers accept Quine's anti-intensionalist position, but they all dispute that semantic content and structure is limited to a theory of stimulus meaning. Each tries different methods to expand the extensionalist base, respectively: truth theory;

a modal realist version of possible world semantics; stereotypes; conceptual role semantics; and a dual-aspect version of conceptual role semantics.

I argue that these theories all fail in the attempt to produce a more powerful extensionalist semantics than Quine's minimalist theory. While each fails for its own reasons, general conclusions may also be drawn. The last three theories share an empirical approach, introducing more content and structure into a language via psychology or via causal relations to the external world; in so doing, they produce either too much structure (conceptual role semantics) or too little (Putnam). The first two share a formal approach, introducing more structure via the imposition of an abstract grammar, which is only later assigned empirical content by a theory of interpretation; but extensionalist limitations on the kind of evidence available to the theory of interpretation force such grammars into the untenable position of producing semantic relations out of weaker ones. The result is that neither of the general approaches open to the extensionalist is viable.

Finally, this result has the following bearing upon the intensionalist-extensionalist debate: the principal case against intensionalist semantics remains embodied in Quine's theory of meaning.

Acknowledgments

My appreciation is due to my dissertation advisor, Professor Jerrold Katz, for guidance and enlightenment; Professor Richard Mendelsohn for counsel and constructive criticism; and the whole Graduate Faculty in Philosophy for an education. They have my sincerest thanks.

Finally, I thank my wife, Melissa Mizel, for her patience and encouragement and also for reading the pages that follow.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION: QUINE'S CASE AGAINST INTENSIONALISM

Traditionally the theory of meaning has proceeded by explaining such matters as meaningfulness and synonymy in terms of meanings. Meaningful expressions have associated with them an abstract object, called variously a meaning, a sense or an intension, which determines both their logical structure and their reference; ambiguous expressions have more than one such object associated with them; and meaningless expressions have no such object associated with them. Synonymous expressions have the same object associated with them. This approach to semantics may be labeled intensionalism, and a contrary approach that denies the existence or usefulness of such objects may be called extensionalism. One of the reasons Frege allowed senses into his ontology was to maintain the truth-preserving substitution of coreferential expressions. A way of avoiding the admission of senses is to strengthen the condition upon the principle of substitutivity. To do that, however, is to admit intensionalism through the back door. So extensionalism involves the rejection of both intensional entities and intensional logics, as we may call those that install a stronger principle of substitutivity.

Quine is an extensionalist; he is also a sceptic about

meaning. One may be an extensionalist and yet quite cheery about the prospects for a theory of meaning. Quine's scepticism about meaning results not from his extensionalism alone, but from a combination of two factors: his repudiation of intensionalist semantics, and the dim view he takes of the yield of semantics without intensions. It is upon the second of these two factors that this dissertation will concentrate. Extensionalism, according to Quine, is inadequate to the task of constructing a robust theory of meaning. Of course, he takes this as no indictment of extensionalism, of whose sound footing he has no doubt. Quine is content to get along without a rich theory of meaning and make do with the minimalist semantics that he develops in Word and Object.[1]

While he may be satisfied with this state of affairs, Donald Davidson, David Lewis, Hilary Putnam, Gilbert Harman and Hartry Field are not. This is not to suggest that they embrace intensionalism, for they share the fidelity to extensional logic and the aversion to intensional entities that together constitute Quine's extensionalism. However, they are, albeit for separate reasons, more sanguine in their outlook than Quine. Holding that extensionalism can yield more than he allows, each of them seeks to forge a richer theory of meaning. To produce this greater yield Davidson avails himself of the formal resources of Tarskian truth theory. Lewis develops a realist approach to possible

---

[1](Cambridge, Mass.: MIT Press, 1960).

worlds. Putnam employs a causal theory of reference. Harman looks to the resources of a more powerful psychological theory than the stimulus-response psychology adopted by Quine. Field seeks to utilize several of these resources.

I shall argue that the attempts of Davidson, Lewis, Putnam, Harman and Field to construct a viable theory of meaning all fail. If I am right in thinking that the resources upon which they draw--a fuller exploitation of the available logical tools, modal realism, a less restrictive physicalism, causal connections inside and outside the head[2]--cover the range of possibilities open to the extensionalist, then their failure means that Quine was correct in his pessimism about meaning and makes his the only position open to the extensionalist semantic theorist. This result is not of interest to extensionalists alone, however. For it would follow that if Quine is right in his criticism of intensionalism, then scepticism about meaning is the only position open to anyone.

In this introductory chapter, I wish to lay out the

---

[2]These are not "extensionalist" resources per se, but resources available to the extensionalist. It may, however, be objected that some of these notions--causality, for instance--are actually intensional, and thus unavailable to the extensionalist. However, what is at issue here is whether semantics invokes special intensional notions that make it discontinuous with the rest of the science. If causality, despite Quine's objections to it, is a notion that is used in physics, then a semantics that employs it will still be on a par with the rest of the sciences. Thus the demand for naturalism would be met, even though it would be a version of naturalism that would give Quine pause.

issues that are at stake. First, I shall take a look at the kinds of considerations about intensionalism that lead Quine to forego it. Second, I shall outline the consequences Quine sees the adoption of extensionalism as having upon the theory of meaning. Third, I shall enumerate those goals and functions of a theory of meaning that Davidson, Lewis, Putnam, Harman and Field maintain can rightfully be reclaimed from the discard pile to which Quine has consigned them.

§1

Philosophers began by studying the nature of the world; then found that they needed to study the mind that represents the world; and then found that they needed to study the language used to represent the world. So what has drawn philosophers to the study of language in the first place is the fact that by means of language we say things that are true and false and talk about things in the world. How it fulfills this function is the primary interest of philosophers of language. Truth and reference come first. Matters of sense and meaning, and philosophers' interest in them, arise only secondarily. If only one could make do with the extensional alone, that would be all to the good. Linguists might interest themselves in meanings and such, but philosophers would not have to bother with such matters so long as the truth-bearing and referential aspects of language could be explained without recourse to them. Theory of meaning would have no claim upon the central place

it has come to occupy in philosophy. The decision to take up the theory of meaning would be a matter of personal taste.

Unfortunately, things do not work out as one might hope. The substitution of extensionally equivalent expressions within contexts created by expressions such as 'believes that . . .', 'it is necessary that . . .' and 'said that . . .' does not preserve truth value. The failure of the substitutivity principle of extensional logic in these contexts provides prima facie evidence of intensionality. Faced with this situation, one may opt for either of two standard strategies.

The first is to strengthen the condition for substitution of expressions within the troublesome contexts beyond mere extensional equivalence. Carnap's version of this strategy, found in his Meaning and Necessity,[3] begins with the elaboration of a set of semantic rules for a language. The new substitutivity principle runs: Expressions of the language are interchangeable in intensional contexts if and only if their equivalence follows from the listed semantical rules of the language. The problem with this approach is that we need some kind of reason for accepting the replacement principle. How does something get to be on the list of semantical rules? And once it has done so, what special privilege thereby accrues

---

[3]2nd ed. (Chicago: University of Chicago Press, 1956), esp. ch. 1.

to it? Quine is correct in claiming that without answers to these questions Carnap's approach fails to be explanatory of analyticity, giving an account only of analyticity-for-language-L.[4]

Another version of this first strategy is to make the condition for substitution not simple extensional equivalence or equivalence derived from semantical rules, but extensional equivalence in all possible worlds. But how do we tell when two expressions are extensionally equivalent in a possible world? This question will be taken up in the course of our examination of the second strategy.

Whereas the first strategy recognizes the existence, side by side with extensional logic, of an intensional logic, or logics--containing the operator 'follows from the semantical rules of L'--the second strategy, Frege's, treats the failure of substitutivity as a mere surface phenomenon. It maintains the unrestricted generality of the laws of extensional logic, such that there is no proliferation of logics with special operators and no fragment of the language in which one set of logical laws is supplanted by another. We should refrain, however, from labeling this strategy 'extensionalist'. Within the troublesome contexts, expressions give up their usual reference and take on new ones. The failure of normally interchangeable expressions

---

[4]See "Two Dogmas of Empiricism," Philosophical Review 60 (1951); reprinted in From a Logical Point of View, 2nd ed., (Cambridge, Mass.: Harvard University Press, 1961), esp. pp. 32-37.

to be interchangeable within these contexts is no violation of Leibniz's law, for these expressions are no longer extensionally equivalent. Thus extensional logic is preserved, but at a significant cost: swelling one's ontology to include intensional entities--senses, propositions, and so forth--as the objects of reference of expressions embedded in propositional-attitude contexts.

Presented with a choice between logical inflation and ontological inflation, Quine casts a blank ballot, abjuring both intensional logic and intensional entities. First, what is wrong with intensional entities? Is it just that they are abstract and immaterial objects that a good empiricist cannot countenance? The answer to this question must be no, since Quine is prepared to accept the ontology of set theory, and sets are abstract objects no matter how concrete and particular their members may be. It is not abstraction <u>per se</u> with which Quine takes issue. What he objects to is admitting objects that lack clear criteria of individuation, for which reason they cannot be quantified over. This objection applies to senses and propositions, as well as to attributes, properties and relations. Abstract objects of this kind, as opposed to sets (which are identical when their members are identical) and extensions of predicates (which after all are simply sets), may be coextensive and yet different. The way to tell if they are identical or different is to determine whether or not the sentences, closed and open, used to express them are

synonymous. What Quine has argued in "Two Dogmas of Empiricism" is that synonymy cannot be had without intensions. And if intensions cannot be had without synonymy, the circularity is vicious. It follows that synonymy cannot be based on meanings construed as intensional objects, since intensional objects require synonymy for their individuation.

But there is another conception of propositions that Quine deals with in "Two Dogmas," the possible-world conception, according to which propositions are specified set-theoretically, and not epiphrastically, as Quine puts it. They are defined as functions from possible worlds into truth values. So ultimately they are just sets, and are not sets admissible entities? In general, yes. But sets are only as good as what they are sets of. A set of poorly individuated objects is at least as poorly individuated as its members. So if possible worlds lack a criterion of individuation, then no matter how formally correct one's definition of a proposition in terms of them, the proposition has not been adequately individuated.

Quine's objection to this approach then is that the notion of proposition relies upon the notion of possible world. What are we to make of possible worlds? We cannot define them as sets of propositions because that would lead us back to where we started. How then do we individuate them? How do we tell when possible worlds are the same or different? If we are not to be driven into the position of

treating them as primitives,[5] then the only recourse is counterfactual statements. And counterfactuals are notoriously problematic. So, Quine maintains in his review of Identity and Individuation,[6] the possible-world conception of propositions attempts to explain the obscure by means of the more or at least equally obscure. And the same goes for the appeal to possible worlds along the lines of the first strategy.

So much for the considerations that lead Quine to his extensionalist position. My interest, however, lies less in how Quine arrives at this point than in where he goes from there. In the next chapter I shall examine in detail Quine's arguments that little progress is possible. Here I shall simply sketch his conclusions, following which I shall outline the ways in which the extensionalist strategies to be examined in the following chapters deviate from the minimalist standard he sets.

§2

In what ways then does Quine's extensionalism limit the scope of the theory of meaning? What is it that one wants out of a theory of meaning but that can no longer be had? The list of casualties is long. In the first place, contexts in which substitutivity fails remain steadfastly

---

[5]This will, in fact, be Lewis's strategy.

[6]Quine's review of Identity and Individuation, ed. Milton K. Munitz (New York: New York University Press, 1971), appears in Journal of Philosophy 69 (1972): 488-497.

opaque. They are not altogether intractable, a point Quine seeks to demonstrate in "Quantifiers and Propositional Attitudes,"[7] but can never be rendered completely transparent. Semantic structure accessible to quantification is not to be found within those opaque contexts. Thus each such context introduces a new semantic primitive. Consequently, the language is encumbered with an infinite number of primitive predicates.

Furthermore, without intensions there is no analytic-synthetic distinction and no synonymy, and consequently no truths in virtue of meaning. The upshot is holism and indeterminacy. Quine employs a variety of metaphors in talking about theories, languages and belief systems: a woven fabric, a web, a force field. One commonality underlying these metaphorical figures is their elasticity. One may hold their edges fixed, yet away from those edges they are capable of being pushed, pulled, stretched and distorted. In Quine's metaphor, it is sensory experience, evidential input from the physical world, that holds the edges firm. But the question arises what determines the interior topography?

Standardly one may distinguish three classes of necessary truths that inform the structure of a theory or belief system--the logical truths, the semantic truths and the modal truths--each of which is distinguished by the

---

[7]Journal of Philosophy 53 (1956): 177-187; reprinted in The Ways of Paradox and Other Essays, rev. ed. (Cambridge, Mass.: Harvard University Press, 1976), 185-196.

brand of necessity that it supports. Quine's rejection of intensions and non-Humean modalities, however, eliminates the last two of the three classes. The truths they contain lose their privileged status and wind up on a par with the nomological truths, which, despite playing an important role in the structuring of theory, are contingent and revisable. In "Two Dogmas" Quine even appears to call into question the privileged status of the logical truths.[8] But elsewhere he resists taking that last step. All truths then, except perhaps for the logical, are in principle open to revision. As a result of this demotion of the semantic and modal truths, the theory builder lacks the resources to determine uniquely the internal structure of the theory. Methodological considerations are insufficient to make up the difference.

Without senses and propositions, no quantum of meaning will attach individually to the vast majority of sentences; their meaning will attach to them only in function of their place in the whole system (or, where Quine mitigates his holism,[9] in a significant chunk of a system). Likewise, no separable quantum of meaning will be assignable to individual subsentential expressions. Inscrutability of reference results. And so does incommensurability, since there will be no common ground either of meanings or of

---

[8]Pp. 42-43.

[9]As in "On Empirically Equivalent Systems of the World," Erkenntnis 9 (1975): 314.

referents on which to compare one theory with another or one belief system with another. Incommensurability and indeterminacy appear closer to home in the form of problems of interpretation of the speech and propositional attitudes of others, since the best available account of propositional attitudes takes as their object not propositions but linguistic entities. These matters will be taken up in the next chapter which addresses Quine's positive theory of meaning, rather than the negative account I am concerned with here.

§3

It is time now to consider briefly where Davidson, Lewis, Putnam, Harman and Field stand on these matters. For Davidson, meaning must be the semantic component of what humans know when they have learnt a language. He seeks to provide a compositional theory of that semantic component. It is because of the importance he assigns to compositionality and learnability that Davidson must reject Quine's treatment of opaque contexts, which saddles the theory of meaning with an infinite set of primitive predicates and thus makes the language unlearnable and insusceptible to a compositional semantics. The theory of propositional attitudes that Davidson offers is designed also to eliminate the language relativity of Quine's analysis.

The key to Davidson's compositional semantics is the casting of meaning in terms of truth conditions. A Tarski-

style truth theory then makes it possible to state the meaning of any sentence and to establish for any two sentences whether they are synonymous, these being two of the functions of a theory of meaning that Quine claims cannot be fulfilled. Truth conditions also supply a common ground upon which two theories may be compared or upon which rival translations of two languages may be compared (although indeterminacy remains because there is always the possibility that two translation schemes will do equally well). Thus propositional-attitude ascription, communication and theory comparison are subject to sound explanation.

Lewis may appear to be the odd man out in the list of extensionalists. His is a possible world semantics. And Quine, we are well aware, rejects possible worlds. However, Lewis's adoption of possible worlds does not involve the adoption of intensional logic or intensional entities. He proposes instead to treat possible worlds and their inhabitants as primitives, a third approach that Quine dismissed out of hand, treating them as concrete, existing entities, different from the actual world and its actual inhabitants only in that they fail to be actual. Given such possibilia, the semantic theorist first constructs grammars that pair expressions with interpretations that assign to them extensions in all possible worlds, not simply in the actual world, as Davidsonian meaning-theories do; and then pairs grammars with natural languages.

On Lewis's account then, while indeterminacy and inscrutability remain, albeit limited in their extent, extension in all possible worlds provides a criterion for synonymy judgments. Propositions are sets of possible worlds, and may thus serve as a language-independent basis for theory comparison. Analyticity, defined as truth at every possible world, is revived. And de re necessity will be admitted, with accessible semantic structure within modal contexts.

Putnam, in departing from the Quinean line, takes a different tack. His account of meaning is built on the idea that matters of reference and meaning are matters not of psychology but of sociology. Reference is determined not by the individual speaker but collectively by the linguistic community to which the speaker belongs. In becoming a member of that community, in learning each expression of the language, the speaker learns "a standardized description of features of the kind that are typical, or 'normal', or at any rate stereotypical."[10] What the speaker has thus grasped in learning the meaning of an expression is a stereotype that serves to pick out the referent of the expression. It may succeed in this task despite being incomplete and possibly inaccurate. The further task of determining the true extension of the expression falls to the linguistic community as a whole and not to the

---

[10]"The Meaning of 'Meaning'," in Mind, Language and Reality, vol. 2, Philosophical Papers (Cambridge: Cambridge University Press, 1975), 230.

individual speaker alone, and is an empirical task.

It is clear from these brief indications that Putnam must depart from Quine on the question of inscrutability. There will be truths in virtue of meaning, i.e., stereotype, but they will not be analytic in that they comport no necessity. While such de dicto necessity is ruled out, there will however be de re necessity. Communication and interpretation of another's speech present no particular problem once it is realized that meanings are "outside the head," that knowing meaning is not a solipsistic act.

The single- and dual-aspect versions of conceptual role theories developed by Harman and Field, respectively, represent the most ambitious of these attempts to restore a complete theory of meaning. Harman takes the ultimate repository of meaning to be thoughts and concepts, elements in an internal system of representation. The meaning of an expression is then to be identified with the conceptual role of the thought or concept that it typically expresses. Thus a specific meaning can be assigned to every expression of the language. Synonymy will be cast as identity, or at least high degree of similarity, of conceptual roles. Propositional attitudes will have an objective correlative. There will be no incommensurability of theories. As for indeterminacy, there will be an objective fact of the matter--conceptual role--that a good translation must capture. Field seeks to achieve the same results, but with truth conditions supplementing conceptual role.

These then are the various extensionalist approaches to the theory of meaning that are to be examined in due course. As we shall see, together they explore the range of options open to the semantic theorist working in Quine's lineage. So if they fail in overcoming the obstacles that Quine sees confronting the theory of meaning, then, short of the development of radically new approaches, the limits of extensionalist semantics are those that Quine has established. Just what those limits are will be the subject of the next chapter.

CHAPTER II

QUINE'S POSITIVE ACCOUNT OF MEANING

In chapter 1, we examined Quine's negative account of meaning, his case against intensionalism. In the present chapter, we turn to his positive account, the semantics he develops, without recourse to intensions and with only acceptable extensionalist tools, in Word and Object. In chapter 2 of that book, Quine, in order to explore the theory of meaning, first recasts matters of meaning in terms of translation. Thus, his pessimism about the traditional semantic enterprise finds its expression in his thesis of the indeterminacy of translation:

> manuals for translating one language into another can be set up in totally divergent ways, all compatible with the totality of speech dispositions, yet incompatible with one another.[1]

In the first section of this chapter I shall lay out Quine's theory of translation based on stimulus meaning. In the following two sections I shall examine the argument for the indeterminacy thesis in order to differentiate it clearly from both the inscrutability of reference and the empirical underdetermination of theory by evidence, while at the same time showing how it is related to each of them. The examination will reveal that, with intensions ruled out of

---

[1]Word and Object, 27.

play, matters of meaning and matters of belief become hopelessly enmeshed, a result that renders impossible a full-blooded theory of either. The fourth section of the chapter addresses the question of the priority of Quine's physicalism and extensionalism, concluding that the physicalist reduction attempted by Quine in chapter 2 of Word and Object is justified only by the prior dismissal of intensions argued for elsewhere, especially in "Two Dogmas of Empiricism." Thus extensionalism is the base of Quine's account of meaning, that upon which his behaviorist version of physicalism rests.

§1

Quine's indeterminacy thesis is about radical translation, that is, translation between two languages that ignores any shared cultural background of the two language communities. Ideally it would be the translation of the language of a hitherto unknown people. Quine envisages a field linguist whose objective data are simply "the forces that he sees impinging on the native's surfaces and the observable behavior, vocal and otherwise, of the native."[2] Out of these data Quine's linguist must construct "the complex of present dispositions to verbal behavior in which speakers of the same language have perforce come to resemble one another,"[3] for such a "complex of present dispositions"

---

[2]Word and Object, 28.

[3]Word and Object, 27.

is what Quine takes language to be.

The development of these dispositions constitutes language learning, which is, according to Quine, a matter of conditioned response to observable stimulations. These stimulations may be either verbal or non-verbal. Some sentences are keyed more particularly to the former, and some more particularly to the latter. Of necessity, the radical translator will begin, like a first-language learner, by correlating native utterances with non-verbal stimulations. Having recorded unelicited utterances, and having tentatively identified certain of these utterances as sentences and certain native expressions as those for assent and dissent, he proceeds to query those sentences in a variety of stimulus situations. In this way Quine's linguist gathers the inductive evidence for translating the native language into his own tongue.

This evidence serves in the construction and assignment of stimulus meanings, which are "the objective reality that the linguist has to probe when he undertakes radical translation."[4] The stimulus meaning of a sentence is an ordered pair of sets of stimulations, <affirmative stimulus meaning, negative stimulus meaning>, whose elements are defined as follows:

> a stimulation σ belongs to the affirmative stimulus meaning of a sentence S for a given speaker if and only if there is a stimulation σ' such that if the speaker were given σ', then were asked S, then were given σ, and then were asked S

---

[4]Word and Object, 39.

> again, he would dissent the first time, and assent the second.[5]

Or more simply, the affirmative stimulus meaning of a sentence for a given speaker is the set of all assent-prompting stimulations. Negative stimulus meaning is defined in the same fashion, but with the terms 'assent' and 'dissent' interchanged. Thus the "stimulus meaning of a sentence for a subject sums up his disposition to assent to or dissent from the sentence in response to present stimulation."[6] And translation proceeds by matching speech dispositions: "there is no justification for collating linguistic meanings, unless in terms of men's dispositions to respond overtly to socially observable stimulations."[7]

Stimulus meaning is not to be equated with any traditional or intuitive notion of meaning. Stimulus meaning is a surrogate. It captures only a sentence's ties to non-verbal stimulation. The more a sentence's use depends on its connections with other sentences, the greater the gap between stimulus meaning and meaning traditionally construed. Such is the case with sentences like 'Ancestral spirits govern the world' and 'The sun is new each day', to which assent or dissent can be elicited without present non-verbal stimulation and which are highly resistant to change

---

[5]Word and Object, 32.

[6]Word and Object, 34.

[7]Word and Object, ix.

in assenting and dissenting behavior.[8] Standing sentences, as Quine calls them, may thus have largely coincident stimulus meanings through sheer paucity of assent- and dissent-prompting stimulations, although they are felt widely to differ in meaning. In the case of occasion sentences--those, such as 'Rabbit', to which assent and dissent must be prompted by present stimulation--the linguist generally fares better. 'Rabbit', the explanation runs, would be uttered or assented to mostly in the presence of rabbit-style stimulations. But even here stimulus meaning achieves only an approximation of meaning. Prior collateral information possessed by the native informant, but unknown to the translator, may be what prompts his assent to, or dissent from, a queried sentence.

The closest a linguist can come to full-fledged meaning is in the case of those occasion sentences whose stimulus meaning does not vary from person to person under the influence of collateral information since they are associated with publicly observable concurrent circumstances, namely, observation sentences. Nevertheless, the linguist only manages to extract meaning <u>plus</u> generally shared collateral information, which cannot be factored out. However, where sentences' use is keyed to non-verbal stimulations, "their stimulus meanings may without fear of

---

[8]My examples are intended to represent the sort of religio-philosophical beliefs that are very theory-laden.

contradiction be said to do full justice to their meanings."[9] So the translation of observation sentences may proceed by inductively establishing the significant approximation of stimulus meanings. However, as sentences grade off toward the non-observational, and intersubjective variability of stimulus meanings widens, stimulus meanings lose their capacity to determine unique translation. Where the disposition to switch from assent to dissent or vice versa in the presence of non-verbal stimulation does not exist, stimulus meaning becomes hollow and translation indeterminate.

All the same, stimulus meanings also provide surrogates for the standard semantic notions of analyticity and synonymy. The linguist can establish "stimulus-analytic" sentences--those which receive assent regardless of prompting stimulation--and pairs of intrasubjectively "stimulus-synonymous" occasion sentences--those pairs which command assent, and dissent, concomitantly. The former cannot be translated on the basis of matching stimulus meaning, because the assent or dissent they elicit may be due as much to collateral information as to "meaning." However, the latter may be translated if the linguist goes bilingual.

Finally, stimulus synonymy is adequate to the task of translating the truth-functional connectives. However, Quine argues, it is insufficient to constrain a unique

---

[9]Word and Object, 42.

translation of the apparatus of first-order predicate logic plus identity. The reasons why and the resulting referential inscrutability will be examined in the next section.

For translation to proceed beyond the yield of stimulus meaning, and into the non-observational zone, requires the introduction of what Quine calls "analytical hypotheses." Reasoning by analogy of syntactic function and projecting the parochial linguistic habits of his own language onto the language of the natives, the linguist:

> segments heard utterances into conveniently short recurrent parts, and thus compiles a list of native "words." Various of these he hypothetically equates to English words and phrases, in such a way as to conform to [the prior yield of translation based solely on stimulus meaning]. Such are his analytical hypotheses.[10]

Conformity to prior yield does not have to be strict; that is to say, while the formulation of analytical hypotheses should seek to respect pairings of sentences on the basis of common stimulus meaning, there is room for give. These initial pairings may be revised if the analytical hypotheses turn out to be simpler and the beliefs of the natives seem more reasonable to us.

Indeterminacy arises because analytical hypotheses "exceed anything implicit in any native's dispositions to speech behavior";[11] this is not so much a defect of analytical hypotheses as the very task they were meant to

---

[10]Word and Object, 68.

[11]Word and Object, 70.

perform, extending "the working limits of translation beyond where independent evidence [in the form of stimulus meanings] can exist."[12] In fact analytical hypotheses are only hypotheses so-called, for they are not capable of independent confirmation or disconfirmation the way translation of 'yes', 'not' or 'and' are. Analytical hypotheses are not determinate functions of linguistic behavior or dispositions. Hence there is no fact of the matter to decide between rival translations.

> [R]ival systems of analytical hypotheses can conform to all speech dispositions within each of the languages concerned and yet dictate, in countless cases, utterly disparate translations; not mere mutual paraphrases, but translations each of which would be excluded by the other system of translation. Two such translations might even be patently contrary in truth value, provided there is no stimulation that would encourage assent to either.[13]

§2

Care needs to be taken to distinguish the indeterminacy thesis from Quine's thesis of the inscrutability of reference--that even if there were a unique translation of all sentences, it would remain unsettled just what the terms occurring in those sentences were true of--for certain of Quine's remarks, incorrectly, so I believe, create the impression that the two theses are more closely related than is actually the case. For example, in "Translational

---

[12]Word and Object, 70.

[13]Word and Object, 73.

Indeterminacy and Private Worlds,"[14] Steven Davis, through misconstruing the import of Quine's discussion of the term 'gavagai', confuses the argument for the inscrutability of terms with that for indeterminacy. Such misreadings were common enough among the early critics of indeterminacy that Quine begins his 1970 article "On the Reasons for Indeterminacy of Translation" by issuing an explicit disclaimer. Nevertheless, in the same place Quine goes on to state:

> There are two ways of pressing the doctrine of indeterminacy of translation to maximize its scope. I can press from above and from below. . . . By pressing from below I mean pressing whatever arguments for indeterminacy of translation can be based on the inscrutability of terms.[15]

Inscrutability arises when the linguist goes below the sentential level and seeks to translate individual terms. Stimulus meaning belongs to each sentence as a complete unit. The observation sentence 'Gavagai' has its own stimulus meaning. It is the same as those of each of the English sentences 'There's a rabbit' and 'There are some undetached rabbit parts'. Both these English sentences are therefore good translations of the native sentence 'Gavagai'. However, we speakers of English naturally prefer the first, and are tempted accordingly to translate the term 'gavagai' by the term 'rabbit'. For this last step to be justified 'gavagai' and 'rabbit' must be coextensional,

---

[14]Philosophical Studies 18 (1967): 38-45.

[15]Journal of Philosophy 67 (1970): 183.

i.e., true of the same objects. Now 'rabbit' is true of rabbits. What is 'gavagai' true of?

To answer this question we need to know where one gavagai stops and the next one starts. That is to say, we need a criterion of individuation. Such a criterion presupposes an independently fixed translation of the apparatus of first-order logic plus identity. We need to know, among other things, when to translate a native locution as the 'is' of identity. In one scheme of translation, wherein we translate 'gavagai' by 'rabbit', a certain native expression may go over into 'is the same as'. However, we could just as well adopt another translation scheme that correlates 'gavagai' with 'undetached rabbit part', in which case that certain native expression would be translated into something like 'belongs with'.[16] On either alternative there is no difference in stimulus meaning of the sentences proposed as translations. So we could equally well adopt either pair of translations.

Underlying this example are the following considerations: In seeking to determine the extension of a predicate we need to be able to identify the objects of which it is true. So we require a characterization of identity. This typically involves two conditions: reflexivity and Leibniz's law, viz., for all suitable replacements of the schematic letter 'F',

---

[16]See "Ontological Relativity," in Ontological Relativity and Other Essays (New York: Columbia University Press, 1969), 33.

$(x)(y)(x = y \rightarrow (Fx \leftrightarrow Fy))$.

But to fill out the second condition we need to know what predicates are "suitable replacements" for 'F'. What is it then that we need to know in order to determine whether a particular predicate is suitable? It seems reasonable to suppose that suitability, is at least in part, a matter of what the predicate is true of. But that is exactly where we started. Translation of predicates and translation of identity and the other elements of the referential apparatus are interwoven. Terms can be translated in different ways without any difference in net output--stimulus meaning or truth value--at the level of sentence translations, provided there is appropriate compensatory adjustment of the treatment of the means of objective reference.

Thus inscrutability of terms has no effect on translation at the sentential level, whereas the indeterminacy doctrine is about difference in net output at sentence level. Accordingly, Quine claims no direct connection between the two theses. Yet, in "On the Reasons for Indeterminacy of Translation," he sees inscrutability as having an indirect bearing on indeterminacy:

> one could imagine with some plausibility that some lengthy nonobservational sentences containing gavagai could be found which would go into English in materially different ways according as gavagai was equated with one or another of the terms 'rabbit', 'rabbit stage', etc.[17]

This would be a mistake. Suppose we have some lengthy, non-

---

[17]P. 182.

observational native sentence s containing the term 'gavagai'. Suppose further that one translation manual pairs 'gavagai' with 'rabbit' and that a second translation manual pairs 'gavagai' with 'rabbit stage'. According to the first manual, s is to be translated by sentence t_1 of the target language; and according to the second manual, s is to be translated by t_2. t_1 is a true sentence; t_2 is false. But if Quine's suggestion is right that the high-level indeterminacy could be traced to the translation of the term 'gavagai', then it would make a difference at the sentential level whether 'gavagai' were translated 'rabbit', 'rabbit stage' or some other way. So we would have grounds to distinguish between these alternative translations; the reference of 'gavagai' would no longer be inscrutable. But suppose, because of the theoreticality of s, there was no way of telling whether the natives take s to be true or false, and so we could not decide on one translation manual over the other; would not the reference of 'gavagai' thus remain inscrutable? Yes, it would; but since 'gavagai' is a term that lies toward the observational end of the spectrum, it makes more sense to think of the indeterminacy as the result of theoreticality, pressing from above, not from below. After all, Quine does speak of "lengthy nonobservational sentences."

Inscrutability and indeterminacy are best kept at different levels. Nevertheless, the two theses are related. Indeterminacy can be expressed as the impossibility of

deciding between rival sets of analytical hypotheses. The first step in generating analytical hypotheses is to cut sentences up into "words." If there existed an independent, fixed translation of terms, i.e., if there were no inscrutability, then there would have to exist a unique way of segmenting sentences into words, and so indeterminacy would be, at the least, severely circumscribed. The indeterminacy would pertain only to those analytical hypotheses, or aspects of them, that dealt with the already-segmented words.

There is also a useful parallel to be drawn between indeterminacy and inscrutability. Inscrutability follows from the doctrine, attributed by Quine to Bentham, that sentences are the primary vehicle of meaning; it results from the attempt to parcel out the quantum of meaning that a sentence has among its component terms. If we take the further step toward holism--even the moderate or relative holism to which Quine subscribes[18]--and take the primary vehicle of meaning to be slices of language that are larger than individual sentences, then it is easy to suspect a parallel difficulty in distributing meaning among the

---

[18]For an explicit retraction of the extreme holism of "Two Dogmas," see, e.g., "On Empirically Equivalent Systems of the World," in which Quine admits "how unrealistic it would be to extend a Duhemian holism to the whole of science, taking all science as the unit that is responsible to observation" (p. 314). While here he is discussing epistemological holism, once this point is combined with the further point, attributed by Quine to Pierce, that meaning is a matter of evidence, the result is a mitigated semantic holism.

individual sentences of the slice. Indeterminacy would thus stand to holism as inscrutability stands to (Quine's reading of) Bentham's theory of fictions.[19]

The parallel is by no means perfect. First, in the indeterminacy case, the linguist benefits from the independently secure translations of observation sentences (which are of no assistance when dealing with term translation); these are not subject to any indeterminacy other than inductive uncertainty, which is not here in question. However, these points of firmness are insufficient to fix the translation of the remaining, non-observational sentences of the language. Second, we lack, in the indeterminacy case, the kind of considerations, such as those above about identity, that would show specifically how the adjustment of the translation of non-observational sentences alone, without tampering with the translation of observation sentences, could compensate for differing translations of a given theoretical sentence. Third, the argument for inscrutability concerns reference whereas indeterminacy is about meaning. This disparity is, however, eliminated, or at least tempered, once we recall that Quine claims that different translations may diverge to the point of differing truth values (extension at sentence level).

---

[19]See "Epistemology Naturalized," in Ontological Relativity and Other Essays, 72.

§3

This parallel with inscrutability is further useful in that it suggests that the case for indeterminacy must be made by pressing from above, working the connection with underdetermination--the thesis that it is inevitable that two physical theories could be empirically equivalent (i.e., agree on all observational evidence, imply all true observation sentences, past, present and future), but be logically incompatible in a non-trivial way (so that the switching of a pair of terms like 'electron' and 'molecule' would not produce underdetermination).[20] The two theses are closely related, as the following passage from "On the Reasons for Indeterminacy of Translation" makes clear:

> Insofar as the truth of a physical theory is underdetermined by observables, the translation of [a] foreigner's physical theory is underdetermined by translation of his observation sentences.[21]

The danger lies in concluding that indeterminacy is nothing more than underdetermination as it applies to semantics. A number of Quine's critics do indeed argue that indeterminacy is nothing over and above underdetermination and so dismiss Quine's special claims for it.[22] Richard Rorty, however, holds a more sophisticated position in

---

[20]See "On Empirically Equivalent Systems of the World": 322-323.

[21]P. 179.

[22]An early example is Jay Rosenberg in "Synonymy and the Epistemology of Linguistics," Inquiry 10 (1967): 405-420.

"Indeterminacy of Translation and of Truth."[23] He describes a dilemma: either indeterminacy is nothing more than underdetermination, or Quine unjustifiably excludes semantics from the theory of nature; and then impales Quine on its second horn. Quine, however, is quite specific about including linguistics as "a part of behavioral science and hence ultimately of physics."[24] Thus semantics suffers from underdetermination as does any science that must generalize inductively from evidence. But that standard underdetermination is not indeterminacy. So let us assume it and then set it aside in order to consider Quine's case of the translation of a physical theory expressed in a foreign language.

The evidence for the physical theory consists of observation sentences; it is relations among these that the theorist seeks to establish and explain. Even on Quine's account, observation sentences have a determinate content. Underdetermination sets in later, when the theorist has to resort to non-observational or theoretical sentences. An unexpected experimental result requires revision of those observation sentences we accept; but it does not specify which theoretical sentences we are to revise or discard; nor does it specify exactly how we are to revise them. This looseness, the flexibility we enjoy to accommodate theory

---

[23]Synthèse 27 (1972): 443-462.

[24]"On the Reasons for Indeterminacy of Translation": 180.

with experience, is underdetermination.

Now, the manual for translating into English a physical theory expressed in a foreign language may itself be looked upon as a theory,[25] one that seeks to establish a system of relations between the sentences of the foreign physical theory and those of our own language. The sentences that the translator seeks to pair up are not confined to the observational; they include the theoretical sentences of the foreign physical theory. Given that "the observational criteria of theoretical terms are commonly so flexible and fragmentary,"[26] so too are the stimulus meanings of theoretical sentences. A good number of theoretical sentences are so well entrenched that their being queried would lead to assent on any occasion. And if a theoretical sentence is not well entrenched, still few stimulations (especially of brief duration) would succeed in converting dissent to assent. A stimulation that did so would constitute an experimentum crucis for the theoretical sentence in question. And Duhem's point about crucial experiments is that they are, at best, hard to come by. The more theoretical a sentence is, the less is acceptance of it directly conditioned by observation sentences. Thus physical theory is underdetermined by the data.

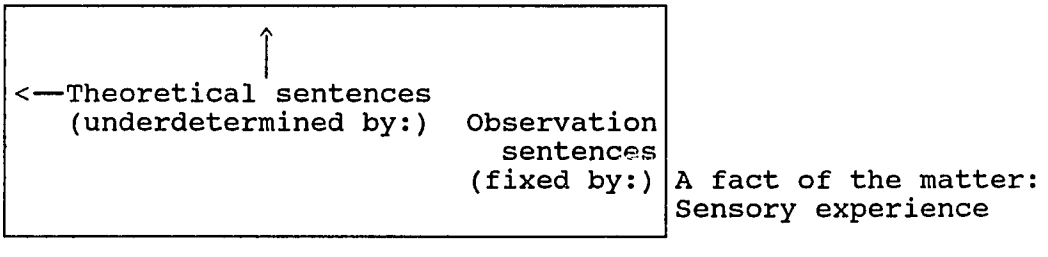
---

[25]See "Reply to Chomsky" in Words and Objections, ed. Donald Davidson and Jaakko Hintikka (Dordrecht: Reidel, 1969), 302-303.

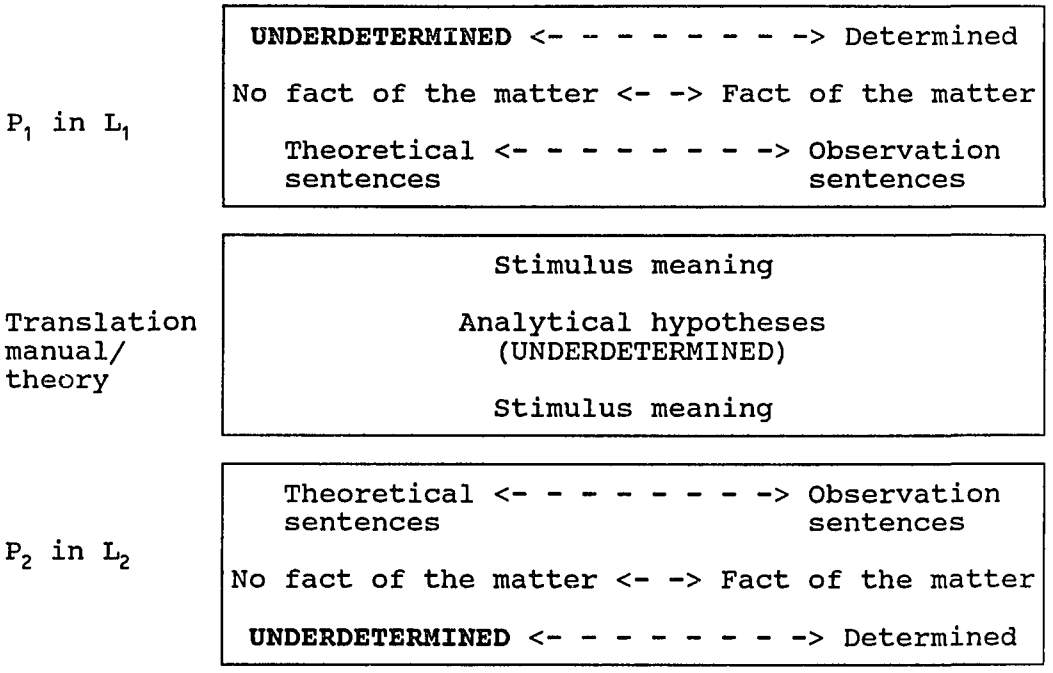
[26]"On the Reasons for the Indeterminacy of Translation": 179.

What makes indeterminacy of translation distinct from straightforward underdetermination is the place that these underdetermined sentences of the physical theory occupy in the translation manual. They are not found in the theoretical part of the manual; they are part of the raw data that the translator has to work with. Where a physicist has objective, determinate data to work with, the translator has to grapple with sentences that are themselves underdetermined, lacking in determinate content.

Let me try to make this point graphically. Think of a physical theory as seeking to establish correlations between observation sentences. Accept that the observational evidence underdetermines the theoretical sentences. There is a fact of the matter, however, about the observational data that the theory has to work with. To adopt Quine's image from "Two Dogmas," the theory is a field of force or a man-made fabric whose edges must square with experience, but whose interior structure is flexible, underdetermined. The more theoretical a sentence is, the further removed from the sensory periphery, then the greater the degree of play, the more it is underdetermined by experience.



Now consider a translation manual, which is to be thought of as a translation theory seeking to establish correlations between the sentences of one physical theory expressed in one language (P_1 in L_1) with the sentences of the appropriate physical theory expressed in another language (P_2 in L_2). The analytical hypotheses of the translation manual will be underdetermined by the translation of the observation sentences of the two theories. This is the underdetermination that affects any empirical theory, semantic or otherwise. However, while the edges of the physical theory are fixed, that is not the case for the translation manual. At the observational end of the spectrum, their content is fixed; but as we move toward the theoretical end, their content is less and less determined, because they suffer from underdetermination. Thus where physical theory has determinate data to work with and is affected by underdetermination only in its theoretical reaches remote from its evidential base, the translation manual encounters underdetermination in its evidential base, for it is trying to pair up not just observation sentences but also theoretical sentences.



Thus the extent of indeterminacy in the translation manual/theory is identical to the extent of underdetermination in the physical theories (in boldface in above figure); but indeterminacy is not merely underdetermination all over again. Translation is subject to underdetermination in just the way any other physical theory is (in parentheses in above figure); but it suffers furthermore from indeterminacy because it is not seeking simply to pair up observation sentences with observation sentences (the purpose of theory construction) but also underdetermined theoretical sentences with underdetermined theoretical sentences. Translation does not simply engender underdetermination the way a regular theory does, but it also encounters underdetermination in what it has to work

with.

To sum up, then, indeterminacy is a consequence of underdetermination.[27] Were there no underdetermination of physical theory, the sentences that a translation manual seeks to pair up would have determinate content; and so there would be no indeterminacy either. Now some may be tempted to say that on this interpretation of indeterminacy, it is no different from underdetermination. But let us get clear on this: indeterminacy is the consequence of the underdetermination of physical theory in general, not of the standard underdetermination from which the translation manual/theory also suffers. A translation theory for two underdetermined physical theories will itself be both underdetermined and indeterminate, while a translation theory for two determined physical theories would be underdetermined but not indeterminate. Thus translation suffers from the effects of underdetermination twice, because its evidential foundation is contaminated, not just its theoretical scaffolding. This is what is unique to translation and semantics, and why semantics is vitiated in a way that other theories are not, why indeterminacy is a serious defect.

Pressing from above in this way is, I think, the best

---

[27]Karl Schick, in "Indeterminacy of Translation" (Journal of Philosophy 69 (1972): 818-832), also concludes that indeterminacy follows directly from underdetermination, if we concede to Quine that physical theory and translation share the same non-verbal evidential base.

case there is to be made for Quine's indeterminacy thesis;[28] it is also the case that makes best sense of those remarks of his that link indeterminacy and underdetermination. It would be nice, Quine acknowledges, to produce an actual example of a pair of rival systems of analytical hypotheses. But duplicating a system of translation from scratch would be a mammoth undertaking.

Harman has tried to come up with something close to an example. In "An Introduction to 'Translation and Meaning,'"[29] he asks us to consider the translation of number theory into set theory, first via Von Neumann's method and then via Zermelo's. Certain sentences, such as 'Three is a member of five', would then turn out true when translated under the first scheme and false when translated under the second. The problem with this example is that it is difficult to conceive of the set-theoretic rendering of number theory as a case of radical translation. If we are dealing with languages at all, we are not dealing with natural languages; we may well ask whether such renderings are intended to capture sameness of meaning, given that the formula 'Three is a member of five' is meaningless in number theory; the notion of stimulus meaning is of doubtful application here; and there is a massive background of

---

[28]One important benefit is that it does not invoke the cumbersome apparatus of stimulus meaning, with the result that a successful challenge to the stimulus-meaning theory does not automatically discredit the indeterminacy thesis.

[29]In Words and Objections, 14-26.

shared information that underlies number theory and set theory.

§4

In any case, writes Quine:

> one has only to reflect on the nature of possible data and methods to appreciate the indeterminacy. Sentences translatable outright, translatable by independent evidence of stimulatory occasions, are sparse and must woefully under-determine the analytical hypotheses on which the translation of all further sentences depends.[30]

We have indeed been reflecting on these "possible data and methods." Given that language is a set of dispositions, which stimulus meaning is designed to capture, translation, having the goal of matching these dispositions, is limited by stimulus meaning. When stimulus meaning peters out, beyond the observational, there is no longer any fact of the matter for translation to be right or wrong about and talk of the meaning of individual sentences becomes meaningless.

Stimulus meaning, as we have seen, is the product of observation of spontaneous native utterances and assenting or dissenting responses to queried sentences. When a native assents to or dissents from a queried sentence, his assent is a function of, on the one hand, what the sentence means and, on the other, what he believes to be true (a combination of sensory information and prior collateral information). In the case of observation sentences, it is safe to assume that what the native believes is true.

---

[30]Word and Object, 72.

Sensory information overshadows prior collateral information, and evolutionary development of the senses, nervous system and brain would select for accurate awareness of the macro-environment; otherwise these natives would not have survived. Beyond the set of observation sentences, sensory information gives way increasingly to prior collateral information that varies from individual to individual, and we can no longer appeal to adaptive value; the capacity for myth and self-deception may have greater adaptive value than the will to truth. Hence, we cannot assume that what the native believes in the case of non-observational sentences is in fact true. So, whereas in the case of the observational we can isolate the contribution of meaning to the native's assenting and dissenting behavior, the twin factors of meaning and belief remain entangled in the non-observational case. In the former case, we can get an independent fix on what the native believes to be true: he believes what is true, so meaning cleaves to the truth, and we have a good empirical theory about what is true. We have no such luck for the rest of the native's language.

Assent and dissent are a function of semantic content and empirical content. The empirical content is a combination of present sensory stimulation and prior collateral information. In the case of observation sentences, the effect of prior collateral information is virtually null. The empirical content thus reduces to present sensory stimulation, and we can determine what that

is. In the case of theoretical sentences, the effect of present sensory stimulation is outweighed by that of prior collateral information, which is the result of personal history and thus widely variable from individual to individual. Hence, we cannot fix the empirical content and isolate the semantic content of theoretical sentences. Whence indeterminacy.

But what is the argument that these data and methods--we may call them "behaviorist," as Quine does in "Indeterminacy of Translation Again"[31]--are the only ones available to the translator? Why is stimulus meaning the only fact of the matter? Why does non-verbal stimulation exhaust the determinate content of our sentences? Why are dispositions to assent and dissent the only ones permissible?

In Word and Object and the two later articles on indeterminacy, these questions receive no explicit answer; Quine takes this behaviorism for granted. This is because the ultimate basis for indeterminacy, the argument for the methodology from which it follows, occurs earlier, in "Two Dogmas of Empiricism." Quine's physicalism is the result of the failure to make sense of analyticity and its circle of related terms:

> If in general I could make satisfactory sense of declaring two expressions to be synonymous, I would be more than pleased to recognize an

---

[31]Journal of Philosophy 84 (1987): 5.

abstract object as their common meaning.[32]

Unless the analytic-synthetic distinction has already been refuted, then intensions would still remain a viable fact of the matter for translations to be right or wrong about. The translator could adopt the principle: a sentence and its translation must express the same proposition, and a predicate and its translation must denote the same concept; or a term and its translation must be synonymous.

My purpose, however, is not to assess the merits of the extensionalist-intensionalist debate. For that we may turn to Jerrold Katz, who, in "The Refutation of Indeterminacy,"[33] argues that in fact the analytic-synthetic distinction and intensions have not been eliminated by the arguments in "Two Dogmas," and thus that indeterminacy remains unproven. The flaw he perceives in the case against analyticity is that it overlooks the contribution of generative grammar. The criterion of scientific respectability employed by Quine in "Two Dogmas" is that all concepts must be capable of non-circular definition. Analyticity fails this criterion. But, argues Katz, in mathematics and the sciences (and, for that matter, in syntax), there are many concepts that would be excluded by such a criterion. Yet they are deemed acceptable,

---

[32]"Facts of the Matter," in American Philosophy from Edwards to Quine, ed. Robert W. Shahan and Kenneth R. Merrill (Norman, Okla.: University of Oklahoma Press, 1977), 193.

[33]Journal of Philosophy 85 (1988): 227-252.

because they are subject to theoretical definition, which violates the non-circularity condition. The terms needing definition receive it in the context of the theory of which they form part; it is not expected that before a term can be accepted into a scientific theory it be given a theory-independent definition. To hold semantics hostage to such a demand is unjustified. And generative grammar provides the requisite theoretical definition.

Whatever the merits of such intensionalist criticisms, my primary goal remains unaffected. That goal is to assess the viability of Quine's semantic lineage. To that end, I assume throughout what follows that Quine is right about intensionalist semantics. All the philosophers whose theories of meaning are to be examined in the chapters that follow eschew intensions. Davidson and Lewis are explicit in their espousal of extensionalism; they are neutral on the physicalism issue. The others take extensionalism for granted and, furthermore, adopt physicalism. Thus each of them belongs to the Quinean lineage. But all agree that Quine has overreached in his attack on the theory of meaning and that he is too pessimistic about the prospects of reconstructing the kind of theory of meaning that intensionalists were under the illusion they had created. Each sees a point where Quine has been overzealous in his methodological purity and overlooked an available resource that did not violate his extensionalist strictures. So I now turn to the first of these post-Quinean theorists.

CHAPTER III

DAVIDSON AND TRUTH THEORY

The indeterminacy thesis serves to focus attention on the problem of isolating the separate contributions that meaning and belief make toward linguistic behavior. To the extent that these elements remain undifferentiated the progress of Quine's radical translator is impeded. He is faced with a sealed circle. As Davidson puts it:

> We do not know what someone means unless we know what he believes; we do not know what someone believes unless we know what he means.[1]

It is this circle that Davidson proposes to break into. He holds that a solution to Quine's problem of separating meaning and belief is possible because we can do for all sentences what Quine maintains can be done only for observation sentences, namely collapse belief onto truth. The solution is to be found in the realization that we must take the natives' beliefs to be true, on the whole. If Davidson is right in this matter, then meaning can be isolated for all sentences, and a robust semantics, built upon a new use of Tarski's work on truth, can be constructed that will serve as a theory of radical interpretation, which

---

[1]"Truth and Meaning," in Synthèse 17 (1967); reprinted in Inquiries into Truth and Interpretation (Oxford: Clarendon Press, 1984), 27.

will at least brighten the prospects of solving some of the thornier meaning-related problems, such as theory comparison, propositional-attitude ascription and translation.

In this chapter I first examine the series of articles in which Davidson develops his theory of meaning, with a view to laying out his semantic program and its motivations. Second, I compare his theory of meaning with Quine's, making explicit the points on which he agrees with Quine and those where he departs from Quine. It will be clear in all this that despite the departures, Davidson follows the main lines of Quine's extensionalist approach, eschewing intensions and cleaving to first-order logic.

Then I critically examine three key areas in Davidson's program: truth conditions and Tarskian T-sentences; learnability and compositionality; and the relationship between radical interpretation and radical translation. We shall see that the program runs into difficulties on all three fronts. These difficulties are sufficient to undermine Davidson's proposed solution to Quine's problem. Davidson fails to produce a theory of meaning in any way richer than Quine's. There are two reasons for the failure of Davidson's semantic program: it doesn't succeed on its own terms as a theory of meaning, and whatever it achieves, it fails to yield a solution to Quine's problem.

§1

We first encounter Davidson's theory of meaning, still in its developmental stages, in "Theories of Meaning and Learnable Languages."[2] This early article examines the prior restrictions that human psychology imposes upon theories of meaning, in particular those that result from the Chomskian observation that in acquiring a natural language, humans develop an infinite aptitude, the capacity to produce and understand novel sentences, in a finite time and by finite means. That humans, possessors of a specific intellectual endowment and set of learning mechanisms, can learn natural languages tells us something about what languages must be like, something that has consequences for the theory of meaning. Coupled with certain empirical assumptions,[3] one consequence is that to learn a natural language is to form a rule-governed competence. As a further result, only a compositional theory, one that treats "the meaning of each sentence as a function of a finite number of features of the sentence,"[4] can account for the learnability of natural languages. If a language is

---

[2]In Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science, ed. Yehoshua Bar-Hillel (Amsterdam: North-Holland, 1965); reprinted in Inquiries into Truth and Interpretation, 3-15.

[3]Among them Davidson includes the following: "that we do not at some point suddenly acquire an ability to intuit the meanings of sentences on no rule at all; that each new item of vocabulary, or new grammatical rule, takes some finite time to be learned" ("Theories of Meaning and Learnable Languages," 9).

[4]"Theories of Meaning and Learnable Languages," 8.

learnable, "it must be possible to give a constructive account of the meaning of the sentences in [the] language."[5] Spelled out this means:

> we must be able to define a predicate of expressions, based solely on their formal properties, that picks out the class of meaningful expressions (sentences) . . . [as well as] to specify, in a way that depends effectively and solely on formal considerations, what every sentence means.[6]

Such then is the challenge that Davidson sees the linguist as facing, to furnish a set of compositional rules that define meaningfulness and specify sentence meanings. Davidson's specific proposal about how to meet it is that a theory of meaning take the form of a recursive Tarskian truth definition. Such a theory would satisfy the conditions for learnability given above, yielding an effective procedure for determining the meaning of every sentence of a given language. But it is not until "Truth and Meaning" that Davidson takes up this initial suggestion and develops it.

The starting-point of "Truth and Meaning" is the same as the earlier article's: "a satisfactory theory of meaning must give an account of how the meanings of sentences depend upon the meanings of words."[7] The rationale is also the same: we must explain learnability, "the fact that, on mastering a finite vocabulary and a finitely stated set of

---

[5]"Theories of Meaning and Learnable Languages," 3.

[6]"Theories of Meaning and Learnable Languages," 8.

[7]"Truth and Meaning," 17.

rules, we are prepared to produce and to understand any of a potential infinitude of sentences."[8]

The task Davidson sets himself in "Truth and Meaning" is to lay out more fully what such a theory of meaning must look like, the requirements it must meet. The article begins by reviewing a number of failed attempts: the theory of meaning construed as (a) a theory of reference; (b) a theory of meanings; and (c) a syntax plus dictionary. From this preliminary survey, Davidson draws certain morals.

The attraction of approach (a) lies in the success of the theory of reference at furnishing an effective procedure that gives the reference of all expressions in terms of the reference of their constituent parts. Identifying meaning and reference would, if all went well, furnish an effective procedure that gives the meaning of all expressions in terms of the meaning of their constituent parts. But all does not go well, for this maneuver results in the attribution of synonymy to sentences that have the same truth value. So semantic compositionality is not to be borrowed directly from referential compositionality. Despite the failure of this approach, Davidson salvages from it the idea that a viable theory of meaning might nevertheless be modeled on the theory of reference.

But simply invoking meanings in place of references--approach (b)--will not do; Davidson accuses intensional approaches of vacuity; the lack of explanatory power

---

[8]"Truth and Meaning," 17.

attributable to meanings leads him to dispense with meanings as entities. If compositionality cannot be borrowed from the theory of reference, it cannot be borrowed from syntax either. Approach (c) takes a recursive syntax for a language and adds a dictionary assigning a meaning to each syntactical primitive. Such stratagems run into Mates's problem.[9] If two expressions have the same meaning ascribed to them by the dictionary, then the substitution in a sentence of one for the other should preserve truth value. But Mates has pointed out that such substitution in propositional-attitude sentences (whose syntax is clearly understood) fails to be truth-preserving: Suppose that 'D' and 'D'' abbreviate synonymous sentences; then consider the true statement:

(1) Nobody doubts that whoever believes that D, believes that D,

and the statement that results from substituting 'D'' for the second occurrence of 'D':

(2) Nobody doubts that whoever believes that D, believes that D'.

If the substitution principle holds, then (2) should also be true; but it isn't.

A further moral Davidson draws from the theory of reference is the need for holism. He takes Frege's dictum

---

[9]"Synonymity," in University of California Publications in Philosophy 25 (1950); reprinted in Semantics and the Philosophy of Language, ed. Leonard Linsky (Urbana: University of Illinois Press, 1952), 111-136.

about words, that only in the context of a sentence do they have meaning, and applies it to sentences, concluding that only in the context of a language do sentences have meaning. As a result, for Davidson, to give the meaning of any expression involves giving the meaning of every expression, that is to say, to give the meaning of an expression requires first developing a theory of meaning for the whole language: "to interpret a particular utterance it is necessary to construct a comprehensive theory for the interpretation of a potential infinity of utterances."[10]

Some of these morals are already familiar, and in any case they do not, individually or collectively, constitute an argument. Rather they serve to put us in a frame of mind receptive to Davidson's positive proposal: that a theory of meaning for a natural language L should take the form of a recursive definition of truth-in-L. A semantics for, say, English should entail all sentences of the form:

(T) s is true if and only if p,

where 's' is replaced by a structural description of an English sentence and 'p' by that self-same sentence. (This is a special case, where the object language is contained in, in fact, identical to, the metalanguage. We shall take up the general case later.) It should, in other words, meet

---

[10]"Belief and the Basis of Meaning," in Synthèse 27 (1974); reprinted in Inquiries into Truth and Interpretation, 148.

the requirements of Tarski's Convention T.[11] The recourse to a theory of truth as a theory of meaning is justified in the following manner: a Tarski-style truth definition yields a set of necessary and sufficient conditions for the truth of every sentence in the language. And knowing a sentence's truth conditions, maintains Davidson, is one sound construal of knowing its meaning. One's willingness to take this step from truth conditions to meaning depends on one's despairing of any account of meaning stronger than truth conditions. But given that our investigation proceeds from the postulated starting-point of Quine's minimalist semantics, such despair is for the present assumed.

A theory of this kind will be compositional, holistic, extensional and, unless an alternative to a Tarskian truth definition is found, recursively axiomatizable.

> The work of the theory is in relating the known truth conditions of each sentence to those aspects ('words') of the sentence that recur in other sentences, and can be assigned identical roles in other sentences.[12]

As for semantic primitives, the theory has nothing to say about their meaning:

> the task was to give the meaning of all expressions in a certain set on the basis of the parts; it was not in the bargain to give the meanings of the atomic parts.[13]

---

[11]See Alfred Tarski, "The Concept of Truth in Formalized Languages," in Logic, Semantics, Metamathematics, trans. J. H. Woodger (Oxford: Clarendon Press, 1956), 152-278; esp. 187-189.

[12]"Truth and Meaning," 25.

[13]"Truth and Meaning," 451.

Davidson simply places a requirement upon them, that they be finite in number in order for the language constituted by them to be learnable.[14]

In "Semantics for Natural Language,"[15] Davidson returns to the same subject, making some modifications and giving more weight to certain points made in the earlier articles. The learnability condition remains key, but it is no longer cast in terms of an actually infinite capacity.

> Even if there is a practical constraint on the length of the sentences a person can send and receive with understanding, a satisfactory semantics needs to explain the contribution of repeatable features to the meaning of sentences in which they occur.[16]

A theory of meaning must still be compositional and recursive.

> An acceptable theory should, as we have said, account for the meaning (or conditions of truth) of every sentence by analyzing it as composed, in truth-relevant ways, of elements drawn from a finite stock.[17]

It must also be explanatory in a non-circular way, and not make illicit appeal to one set of semantic concepts in order to explain another set of semantic concepts. There should be a prohibition on "the appearance of a semantic term in the statement of the truth conditions of a sentence unless

---

[14]"Theories of Meaning and Learnable Languages," 9.

[15]In Linguaggi nella Società e nella Tecnica (Milan: Edizioni di Comunità, 1970); reprinted in Inquiries into Truth and Interpretation, 55-64.

[16]"Semantics for Natural Language," 55.

[17]"Semantics for Natural Language," 56.

that sentence already contains the semantic term (or a translation of it)."[18] Finally the theory should "provide a method for deciding, given an arbitrary sentence, what the meaning is"[19] and be empirical and testable. This it achieves by generating as theorems the required biconditionals.

§2

Such is the place and function of a truth theory in Davidson's account of meaning. Let us turn now to a comparison of the account of meaning presented in these articles with Quine's account. This section contains a summary of what Davidson promises to deliver that Quine cannot. The primary innovations stem from the appeal to truth conditions and compositionality. Each sentence of the language whose semantics is being constructed will have a meaning assigned to it, namely its truth conditions, which, in turn, supply a criterion by which to judge the synonymy of pairs of sentences. In addition, for each sentence of the language, there will be a sentence, the T-sentence associated with it, that states its meaning. T-sentences will be generated by the semantics for a given language. They may thus be said to be truths that hold in virtue of meaning; however, they are not analytic but empirical truths following from the meaning-theory.

---

[18]"Semantics for Natural Language," 57.

[19]"Semantics for Natural Language," 56.

The theory of meaning will furnish a theory of radical interpretation, the knowledge that would enable one to say what a speaker means by an utterance upon a particular occasion: the knowledge of a truth theory, that it is a truth theory and that it satisfies the further criteria outlined in the previous section. Such knowledge will put an end to problems having to do with theory comparison and communication. The idea of separate, incommensurable conceptual schemes gives way to the radical interpreter. Truth conditions provide a fact of the matter for cross-theoretic interpretation, as well as for translation between languages that appear to share no conceptual commonality. The theory of radical interpretation makes up for the shortcomings of radical translation.

Compositionality entails more semantics, that is to say, more semantic structure. For Quine, the opacity of propositional-attitude ascriptions is ineliminable; within every propositional-attitude sentence lies a core that cannot be analyzed into primitive components, the object of the propositional attitude. Each of these constitutes a new and intractable primitive; their number multiplies to infinity. Moreover, on Quine's account, every propositional attitude is to be construed as essentially a relation between a person and a sentence, whether open or closed; the language relativity of propositional-attitude ascriptions is unavoidable.[20]

---

[20]See Quine, "Quantifiers and Propositional Attitudes."

On Davidson's account, all sentences, including those ascribing propositional attitudes, are built up out of a finite stock of primitives by the application of a finite number of rules. While it is not my task to enter into the question of the merits of Davidson's paratactic account of propositional attitudes,[21] the adoption of a compositional semantics does allow for greater optimism regarding a solution in this area, if only because of the absolute hopelessness of the problem given Quine's approach.

Notwithstanding these departures of Davidson's, there remain important points of agreement between him and Quine. No necessity, either de re or de dicto, is invoked. Davidson, like Quine, restricts his logical apparatus to first-order predicate logic. He rejects any recourse to intensions. There is no analytic-synthetic distinction. For the most part, he assumes the correctness of Quine's arguments, and proceeds without any appeal to these notions. In "Truth and Meaning," he also invokes, as we have seen, Mates's argument against intensionalism. Synonymy is not a strong enough condition to support substitution salva veritate in opaque contexts. If that is correct, then the intensionalist definition of analyticity--two sentences are analytic if and only if the substitution of synonyms for synonyms results in logical truths--cannot be relied upon if substitution may result in change in truth value.

---

[21]See "On Saying That," in Synthèse 19 (1968-69); reprinted in Inquiries into Truth and Interpretation, 93-108.

Davidson's adoption of Tarski's approach entails holism. The power of a theory of truth lies in its total yield, the generation of the complete set of T-sentences, which fixes the extension of the truth predicate. Likewise, the semantics for a natural language functions holistically. What it provides is the relationships of interdependence that hold between the sentences that contain the same recurrent features or "words." The meanings of sentences are not given individually, but all together, as they are assigned a place in the total language. A single T-sentence, while it gives the truth conditions, does not, in isolation, give the meaning of the sentence named on its left-hand side. The T-sentence conveys meaning only insofar as it is a product of the total semantics for the language.

Although its scope is restricted, indeterminacy remains, in that there will be no single theory of meaning or interpretation that is superior to its competitors; more than one theory may satisfy the constraints. Revisability in principle of all beliefs, the logical truths excepted, is still the case. Davidson is, however, clearer than Quine on the exemption of logic. First-order logic must hold good for all languages, for it provides the minimal basis upon which any attribution of rationality or language use is justified.

Reference remains inscrutable. That is to say, the theory operates at the sentential level. In fact, reference is reduced to the status of a theoretical construct, part of

the apparatus posited to implement a theory of truth "without needing independent confirmation of empirical basis."[22] Truth itself is treated as a property and not as a word-world relation; truth almost ceases to be a semantic phenomenon, in the sense of being about the world.

§3

What then are we to make of Davidson's proposals and promises? He has presented, not a theory of meaning per se, but a program for the construction of theories of meaning, a set of constraints that a theory of meaning for a natural language should satisfy: "what is being recommended is not a particular theory, but a criterion of theories."[23]

The first of these constraints has to do with truth conditions. To give the meaning of a sentence is, according to Davidson, to give its truth conditions. However, it does not follow from this that meaning is simply to be identified with truth conditions. The divergence between the two, as evidenced by Frege's 'Hesperus = Phosphorus' puzzle, puts the lie to such an interpretation of Davidson. Frege developed his theory of meaning in response to the failure of purely referential theories of meaning to account for certain puzzles concerning identity sentences and sentences ascribing propositional attitudes. Referential theories of meaning cannot account for the differing "cognitive value"

---

[22]"Reality without Reference," in Dialectica 31 (1977); reprinted in Inquiries into Truth and Interpretation, 222.

[23]"Semantics for Natural Language," 183.

of sentences of the form:

(3) Hesperus = Hesperus, and

(4) Hesperus = Phosphorus.

Since (4) is true, 'Hesperus' and 'Phosphorus' are coreferential, and so, on the referential account, (3) and (4) must have the same meaning. Yet, (3) is analytic and known a priori, while (4) is synthetic and the discovery of its truth constitutes a genuine addition to our astronomical knowledge. So there must be more to meaning than simple extension, and more to the theory of meaning than simply the construction of Tarskian truth theories.

This first constraint should therefore be construed in a less simplistic manner: A semantics for a natural language must give the truth conditions for every sentence of the language, by generating the T-sentences that express those truth conditions. Simply entailing the set of all T-sentences is not enough, however. For the set of all T-sentences implies all its own members, but alone tells us little about meaning. The role of T-sentences in Davidson's account of meaning is evidentiary. They do not form part of the theoretical apparatus of the semantics for a given language. Rather, being directly available to speakers of a natural language, they serve as that against which the semantics for the language is tested. Their role is analogous to that of observation sentences in the testing of a theory that invokes posited entities that are either

unobserved or indirectly observed.[24]

What is required is a theory that, aside from its logical axioms, is finitely axiomatizable and entails the T-sentences. Nothing in Davidson's explicit requirements mandates a Tarski-style theory. However, the merit of a Tarski-style truth theory is that it is already in place. What remains to be done is to extend its domain of application from formal languages to natural languages. But here again there is a risk of oversimplification.

The gap separating Tarski and Davidson is much deeper and more fundamental than has just been suggested. In fact, Davidson himself acknowledges the divergence. To deny it would be a serious violation of Tarski, for what Davidson is up to is nothing less than turning Tarski on his head. Tarski's truth theory is a theory that, taking translation for granted, yields a definition of a predicate, 'is T', that fulfills the requirements that one expects of the predicate 'is true'. The Davidsonian "truth theory" takes truth for granted and expects to yield translation or meaning. So it is inaccurate to think of what Davidson is doing as simply an extension of Tarski.

In fact, calling both theories 'truth theories' is something of a misnomer, akin to identifying an electric motor with a generator. The two devices work on the same physical principles, the former converting electrical energy

---

[24]See "The Structure and Content of Truth," Journal of Philosophy 87 (1990): 300.

into mechanical energy and the latter doing the reverse. However, just because you have an electric motor, it does not follow that you can simply reverse it and thereby create a generator. The electric motor requires a battery or some similar source of power. The generator requires a mechanical power source. Just because you have one power source does not mean you have the other. Tarski assumes translation and draws out a theory of truth; Davidson cannot blithely assume truth and extract a translation relation.

What the Davidsonian and Tarskian enterprises have in common is a mechanism of finitely axiomatizable rules that generates T-sentences. Both theories employ a metalanguage and an object language. In Tarski's account, it is assumed that we are able to translate between object language and metalanguage. This assumption is allowable because Tarski's enterprise is not one of "radical" truth definition, in the way that Quine's is one of radical translation and Davidson's one of radical interpretation. By pairing up the infinite number of synonymous sentences from the two languages, Tarski succeeds in defining a truth predicate for the object language. Thus the infinite pairing of sentences furnishes a definition of truth in the object language. The predicate 'is true in the object language' will have to include all and only those pairs. The requirement that 'p' be the translation into the metalanguage of the object-language sentence '<u>s</u>' provides a sufficient constraint upon the theory to specify a truth predicate.

The use to which Davidson puts the Tarskian apparatus is altogether different. This time, an understanding of truth is intended to yield an infinite pairing of sentences with their meanings. Speakers of the object language know the truth conditions of sentences of that language. Thus, they can substitute the right pairs of sentences for the occurrences of '_s_' and 'p' in T-sentences. However, Davidson himself acknowledges that T-sentences alone do not constitute enough of a constraint to produce a theory of meaning. Further empirical and formal constraints must be forthcoming. And these constraints cannot be had simply by looking to Tarski. They must be sufficient to produce meaning without being guilty of the fallacy of _petitio principii_.

§4

A second constraint is that an adequate theory of meaning must explain how the meaning of a sentence depends on the meaning of the words--or repeatable, recurrent features--out of which it is composed. An essential task of a semantic theory, according to Davidson, is to explain the compositional structure of natural language. Otherwise, there is no explaining how those who have learnt a natural language are capable of producing and understanding an indefinite number of novel sentences. The compositionality constraint thus turns upon the argument from learnability.

While learnability may seem, ever since Chomsky, rather straightforward and familiar to us, there are a number of

questions that need to be raised in connection with it. First, there is some question whether human linguistic capacities are actually infinite, or infinite in an important sense. We have seen that Davidson addresses this issue by backing off on the claim of infinitude in "Semantics for Natural Language." He maintains, however, that his point remains unaffected. Humans have an indefinite, if not infinite, capacity to produce and understand novel sentences; and this is sufficient to generate the demand for compositionality.

But even this weaker stance is troublesome. Davidson is reasoning from a fact about human psychological limitations to a conclusion about the nature of language. If our model of the human learning mechanism is limited, does it follow that what is learnt is thereby limited, as well? If our theory of meaning is constructive, does the language itself have to be constructive? We can distinguish at least three levels at which compositionality may be ascribed: the language-learning mechanism; the theory of the language, possession of which enables one to use the language; and the language itself.

Assuming an argument for compositionality at the first, language-learning level, there remains a need for an argument from there to the third level. Davidson fails to separate out the three levels because he views the ontological status of language as psychological. A language is not something over and above its psychological

instantiation, a mechanism in the speaker/interpreter that corresponds to the theory of the language.[25] The theory of a language is a theory of language understanding, of what the speaker knows in order to be able to produce linguistic utterances and to interpret the speech of others. Davidson considers it a theory of competence, but just where he draws the line between competence and performance is not clear. In either case, it is clear that his argument from learnability remains based in human psychology and unclear whether it can support a claim beyond psychology.

§5

That a theory of meaning should be a theory of language understanding constitutes the third constraint. A Tarski-style theory needs to be complicated by relativization to speaker, time and utterance in order to cope with the presence in natural languages of indexical expressions, which are absent from formal languages. Davidson wants no semantical notions to be introduced on the right-hand side of a T-sentence unless they are already contained on the left-hand side. In the simple case of, say, '"Snow is white" is true if and only if snow is white', there is no problem. But where indexicals occur on the left-hand side, work needs to be done on the right. The rationale for this is the third constraint, that a theory of meaning must also be a theory of radical interpretation, such that possession

---

[25]See "Belief and the Basis of Meaning," 141.

of a theory of meaning for a language is sufficient for interpreting the language. In particular, the possessor of this theory must be able to interpret the object language without recourse to any extra semantical apparatus, such as a theory of denotation. Any such appeal would simply defer the task of interpretation, rather than solve the problem. It also means the loss of empirical testability, for which are needed unadulterated T-sentences that semantic intuition can verify.

Yet deferring the task of interpretation is just what Davidson seems to do, for it appears that his theory makes appeal to a Quinean translation manual, or some equivalent. A Davidsonian truth theory is supposed to entail all T-sentences. When the object language and metalanguage are both, say, English, determining the set of T-sentences seems simple. Any English speaker can do it. But when the object language is not part of the metalanguage, when we are in the situation of the radical interpreter, how are we to determine the set of T-sentences? How do we confirm the truth of the T-sentences for a particular natural language? We must not assume that we can simply translate the foreign language into English. If Davidson is to avoid the charge of making illicit appeal to a theory of translation, of simply translating the object language into a previously understood metalanguage, he must explain the basis for determining T-sentences generally. The practical ease with which we can identify them in the English-language case

disguises a deeper problem. It just may be that a theory of translation is needed to sanction T-sentences.

Davidson's solution draws from the same source as Quine's: the assenting and dissenting behavior of the natives. What this behavior tells us is not stimulus meaning. Rather it tells us what the natives hold true. The concomitant circumstances will then be the truth conditions of the sentences assented to and held true. Here is where we began with Davidson. What is the difference between his appeal to assent and Quine's appeal to it such that the former yields truth conditions for all sentences instead of interesting stimulus meanings for just some? Why does Davidson assume that the natives' beliefs are in general true regardless of their observational status?

His defense is initially methodological. Massive disagreement vitiates the whole enterprise.

> If we cannot find a way to interpret the utterances and other behaviour of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything.[26]

In response to the objection that this argument simply makes a virtue of our limitations as interpreters, converting a truth about our epistemological shortcomings into an objective fact about language, Davidson insists that even an

---

[26]"Radical Interpretation," in Dialectica 27 (1973); reprinted in Inquiries into Truth and Interpretation, 137.

omniscient interpreter would be in the same boat.[27]

But the problem with this defense is that it is directed against a sceptic position, Quine's. Davidson does not realize the full extent of Quine's scepticism, mistaking the import of indeterminacy. In "Reality without Reference," he states that Quine "argues that the totality of evidence available to a hearer determines no unique way of translating one man's words into another's"[28] (emphasis added). Thus Davidson continues to think that indeterminacy is an epistemological and methodological problem, whereas Quine, we have seen, goes much further, maintaining that there is no fact of the matter.[29]

So given this more fundamental scepticism, the claim

---

[27]See "The Method of Truth in Metaphysics," in Studies in the Philosophy of Language, vol. 2, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein (Minneapolis: University of Minnesota Press, 1977); reprinted in Inquiries into Truth and Interpretation, 201.

[28]P. 220.

[29]Moreover, in his most recent treatment of this issue, Davidson seems to ignore the limitations that radical translation and/or interpretation place upon the linguist. In allowing the interpreter to invoke distal stimuli, and not Quine's proximal stimuli, Davidson attributes to the interpreter the ability to pick out that part of the environment that prompts the speaker, the ability to discriminate the cause of the speaker's assent and dissent:

> An interpreter bent on working out a speaker's meanings notes more than what causes assent and dissent; he notes how well placed and equipped the speaker is to observe aspects of his environment, and accordingly gives more weight to some verbal responses than to others ("The Structure and Content of Truth," 321).

But in attributing these abilities to the interpreter, Davidson is equipping the interpreter with information that cannot be assumed in the radical situation.

that if the argument is rejected then truth will turn out to be intra-theoretic and the theory of meaning will go under surely leaves the sceptic unmoved. Moreover, the parochialism Davidson advocates depends upon a particular criterion of rationality, supported by the claim that there is no alternative. He makes this case in "Thought and Talk":[30] Beliefs receive their content through their place in a pattern of beliefs; that pattern of beliefs receives its structure through correlation with the world; if there is no correlation, there is no pattern; and if there is no pattern, there are no beliefs; so no rationality is attributable to the subjects whose speech we seek to interpret.

But if Davidson is right, then the speculations in Descartes's First Meditation would be incoherent. Yet Descartes succeeds in rationally conceiving that he himself (never mind the native; it is not only indeterminacy that begins at home) is massively mistaken, systematically in error. So Davidson's claim that true belief is our only criterion of rationality must be wrong. Even if he wishes to argue that the First Meditation is incoherent, it will take an argument (and criterion of rationality) over and above the impossibility of massive error.

In any case, the claim that shared true beliefs constitute the only criterion of rationality we've got again

---

[30]In Mind and Language, ed. Samuel Guttenplan (Oxford: Oxford University Press, 1975), 7-23; reprinted in Inquiries into Truth and Interpretation, 155-170.

leaves the sceptic unmoved; that we lack a criterion of rationality altogether is a state of affairs that the sceptic can well live with. And as we shall see in chapter VI, Harman offers an alternative criterion of rationality, one based on internal coherence of beliefs and not on their correlation with the external world, which thus challenges the second premise in the above argument. First, however, let us examine Lewis's approach.

CHAPTER IV

LEWIS AND POSSIBLE WORLDS

§1

The stated aim of the present discussion is to examine theories of meaning that belong in the Quinean lineage. Each of the theories under consideration therefore attempts to develop a richer semantics not by overturning Quine but by making a small revision in his approach that will suffice to produce the further yield. With this in mind, the first question that needs to be addressed in the present chapter on David Lewis's possible world semantics is whether it belongs here at all, whether the invocation of possible worlds does not automatically place Lewis's theory beyond the Quinean pale. Yet Lewis's extreme version of modal realism is a result, albeit paradoxical, of his acceptance of Quine's arguments against ontologically weaker conceptions of possible worlds.

Quine treats modal idioms as unanalyzed primitives and modal contexts as ultimately closed. Lewis proposes to open them up. But as soon as this happens, is he not thereby enmeshed in the intensional? Extensionality fails in modal contexts. So he must, it appears, go the route of either intensional logic or intensional entities; and neither route is acceptable to Quine, for the reasons adduced in the

introductory chapter. Lewis, however, seeks to steer a course that satisfactorily overcomes Quine's objections to modality and yields an account of analyticity, synonymy and the rest of the semantic notions that Quine argued must be forsworn. And it is this course that leads him to his version of modal realism.

To begin with, Lewis repudiates quantified modal logic (but not its results), replacing it with "counterpart theory." Counterpart theory is designed to capture modality without the introduction of special-purpose modal operators. It makes do with first-order logic plus identity. Then after the extension of the domain of quantification and the introduction of four predicates--'is a possible world', 'is in a possible world', 'is actual' and 'is a counterpart'--and eight postulates, each sentence of modal logic is to be translated into a sentence of the new theory.[1]

Here is how the theory can deal with the apparent failure of substitutivity within modal contexts (Quine's general objection to quantified modal logic). The problem is the following:

(1) $(x)\ [x = 9 \rightarrow (\Box\, x > 5 \leftrightarrow \Box\, 9 > 5)]$

is false because

(2) The number of planets = 9

and

---

[1]See "Counterpart Theory and Quantified Modal Logic," Journal of Philosophy 65 (1968); reprinted in Philosophical Papers, vol. 1 (Oxford: Oxford University Press, 1983), 26-46.

(3) □ 9 > 5

are true, while

(4) □ the number of planets > 5

is false.

But on Lewis's reading of the necessity operator, (2) is modified to

(5) The number of planets at the actual world w_a = 9

and (4) becomes

(6) For all worlds w, at w the number of planets at w_a > 5,

which is either true so long as (2) is true, or possibly deviant. In either case, (3) and (4) do not have differing truth values, and so substitutivity is preserved.

But at what cost? The idea of extending the domain of quantification to include possible worlds and their inhabitants may well give us pause. How can this be good, old-fashioned first-order logic if it is now to quantify over possibilia? Surely that means that the existential quantifier has changed in some important sense, if it is now to range over possible individuals. Or is Lewis giving up the Quinean position that the quantifier is the measure of our ontological commitment? Lewis doesn't need special modal operators because he's surreptitiously changing the old ones!

But this charge is premature. Lewis must, and does, deny that he is introducing a new quantifier. He argues that the extended domain of quantification contains things

that are no different from the kinds of things that we have been used to quantifying over. What then is his new ontology that isn't new after all?

Lewis rejects linguistic and conceptual treatments of modality, for the sort of reason that we have seen Quine give: where they don't lead into error, they are circular. Lewis is no reductionist; he wants no part of the "ersatzism" that would make of possible worlds state-descriptions, maximal consistent sets of sentences, mere set-theoretic constructions or mental entities. He wants no abstract surrogates for concrete possibilia. Possible worlds are what they are, and not another thing.

Lewis's proposal is to treat at face value the existential quantification in the locution: There are many ways things could have been besides the way they actually are. That is to say, there are other worlds like ours, ways our world might have been; and the existential quantifier will range over "ways things could have been" and their parts, namely, possible worlds and their inhabitants. But the quantifier has not changed because these possible worlds are the same kind of thing as our world. They are concrete, existing in the same way as our world, except that our world is the only one that is actual. Possible worlds exist, but they don't <u>actually</u> exist. And the same goes for their inhabitants. Possible worlds and the things that are in them are taken as primitive entities that are not dependent for their existence upon our words or thoughts. (Possible

worlds had better not depend on them because, as we shall see, they are used to define what a language is.)

Thus, *no intensional logics and no intensional* entities. But do Lewis's possibilia run afoul of Quine's demand for clear criteria of individuation? Failure to meet this demand was, we have seen, the principal objection that Quine brought against the conception of propositions and properties, supplied by possible world semantics, according to which a proposition is roughly a function whose domain is possible worlds and whose range is truth values, and properties are functions from possible worlds to sets of entities in those worlds, possible entities. While functions are acceptable in principle, in particular cases they are only as acceptable as their members. And the problem with "unactualized possibles," as Quine emphasizes in "On What There Is"[2] and Word and Object,[3] is that we cannot tell when we have one possible object or two or a dozen; nor can we tell whether a possible object in one world is the same as this possible object (rather than another) in a second world. Lewis agrees with Quine that individuation and identity are of the utmost importance, but disagrees that possible individuals fall short on this score.

Consider the following simplified example. Suppose there are just two objects, X and Y, in world w_0. They have

---

[2]Pp. 3-4.

[3]Pp. 245-246.

exactly six properties each; X has properties x_1-x_6; Y has properties y_1-y_6. Assume there is trans-world identity. In w_1, X and Y swap the first property; in w_2, they swap in addition the second property; and in w_3, they swap in addition the third property.

World	X's properties						Y's properties					
w_0	x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3	y_4	y_5	y_6
w_1	y_1	x_2	x_3	x_4	x_5	x_6	x_1	y_2	y_3	y_4	y_5	y_6
w_2	y_1	y_2	x_3	x_4	x_5	x_6	x_1	x_2	y_3	y_4	y_5	y_6
w_3	y_1	y_2	y_3	x_4	x_5	x_6	x_1	x_2	x_3	y_4	y_5	y_6

Now let's start again from scratch. In w_0, X has properties x_1-x_6 and Y has properties y_1-y_6. In w_a, X and Y swap the sixth property; in w_b, additionally the fifth property; and in w_c, additionally the fourth.

World	X's properties						Y's properties					
w_0	x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3	y_4	y_5	y_6
w_a	x_1	x_2	x_3	x_4	x_5	y_6	y_1	y_2	y_3	y_4	y_5	x_6
w_b	x_1	x_2	x_3	x_4	y_5	y_6	y_1	y_2	y_3	y_4	x_5	x_6
w_c	x_1	x_2	x_3	y_4	y_5	y_6	y_1	y_2	y_3	x_4	x_5	x_6

Now it turns out that w_3 and w_c are the same world, having just two objects, each having the same properties. But we have a conflict in that, according to the first sequence of worlds we considered, X is the thing in w_{3c} that has properties y_1-y_3 and x_4-x_6, but according to the second sequence the thing with those properties is Y. The problem

lies in the transitivity of identity: if X in w_n is identical to X in w_{n+1} and X in w_{n+1} is identical to X in w_{n+2}, then X in w_n is identical to X in w_{n+2}. Couple this with the fact that in going from one world to the next the properties of the identical object are modified, and as we proceed we may encounter an object in a world that more closely resembles X at the actual world than does X at that world.

Lewis's solution to this kind of problem is to deny trans-world identity. According to his counterpart theory no entity belongs to more than one world, either by being entirely in both worlds or by being a trans-world object that is partially in one and partially in the other. Trans-world identity is replaced by relations of comparative similarity, which are intransitive. And problems with comparative similarity between objects in different worlds are no worse than those encountered in dealing with intra-world similarity relations.

Next I turn to one of the theoretical benefits of this purportedly cost-free theory of modality, namely, the account of semantic content. Then, after laying out Lewis's version of possible world semantics, we shall compare the yield of his richer theory with that of Quine's before proceeding to see whether Lewis's theory holds up under scrutiny.

§2

Like Davidson, Lewis ascribes to truth conditions the number one priority in a semantics for a natural language.

And also like Davidson, he recognizes that truth conditions alone are inadequate to the task of capturing meaning. Where Davidson sees the solution as lying in the construction of meaning-theories that pair sentences and their truth-conditions in accordance with certain further constraints--finite axiomatizability, compositionality and interpretation--Lewis envisions for each language a grammar, encompassing both syntax and semantics, that pairs expressions with interpretations in accordance with a set of constraints that is very different from Davidson's. These interpretations are to incorporate not simply truth conditions in the actual world but truth conditions in all possible worlds.

Davidson already recognized the awkwardness that results from a meaning-theory generating as a theorem a T-sentence such as '"Snow is white" is true if and only if grass is green'. Since there are possible worlds where 'Snow is white' and 'Grass is green' differ in truth value, no such unfortunate pairing offers itself in a possible-world approach, thus obviating the further maneuvering into which Davidson is forced.

We shall study the nature of Lewis's "interpretations" shortly. But first let us see how Lewis develops the grammatical mechanism that assigns them to expressions. The task is subdivided several times over. The first subdivision is between formulating abstract grammars, on the one hand, and on the other, determining which of these

grammars accords with the linguistic practice of the population using the given natural language. The latter task is itself a part of the task of radical interpretation, and is rightly to be preceded by the construction of the grammar, which itself subdivides into two parts, syntax and semantics. The semantic component will assign to expressions interpretations that specify truth conditions. But the syntactic component must first yield up the structures upon which the semantic component is to operate.

Both syntax and semantics subdivide into two parts. The grammar as a whole must be humanly usable and hence finite even though it is to specify an infinite number of pairings of expressions and interpretations. So the grammar is constructed by first setting up a finite lexicon of basic expressions assigned a syntactic category and semantic value, and second deploying a set of compositional rules for building complex expressions out of simpler ones and assigning them a syntactic category and semantic value.

Lewis describes an abstract, systematic grammar for a natural language in his book Convention.[4] The output of the grammar is a function that assigns an interpretation to every member of a set of sentences. I shall treat of the two parts of the grammar, syntactic and semantic, in order. Since the size of the domain of the function referred to above is infinite, and the language is humanly usable, the

---

[4](Cambridge, Mass.: Harvard University Press, 1969), esp. ch. 5.

grammar must be finitely specifiable. So the structure of the grammar that Lewis presents is generative. It consists of three components: a lexicon (a finite set of elementary constituents with syntactic category markers); a generative component (a finite set of combining operations); and a representing component (producing verbal expressions, the sentences of L). If the elementary constituents are taken to be expressions already, we may simplify things by dropping the third component out of the grammar. So basically we have one set of rules that assigns a base structure to items drawn from a finite vocabulary and then a set of transformational rules that produce a surface structure out of that base structure.

What of the semantic aspect, the role of a grammar in the assignment of interpretations to sentences of the language? Lewis borrows Katz's 1966 method:[5] Grammars start by assigning interpretations to lexical elements, then perform projection operations upon them to produce interpretations for larger constituents, and finally transfer sentential interpretations to sentences themselves. The point at which semantic and syntactic aspects meet is the lexicon, to which the syntax assigns a grammatical category and the semantics assigns an interpretation upon which a series of compositional operations is performed.

But Lewis's theory is not Katz's. It is in the

---

[5]Philosophy of Language (New York: Harper & Row, 1966), ch. 4.

conception of the nature of the assigned interpretations that Katz and Lewis go their separate ways. In contrast to Katz (in whose work the semantic markers of Philosophy of Language have since given way to senses), Lewis wishes to make do with extensions, although the extension of a sentence is not limited to its truth value in the actual world, but extends to its truth value in every possible world. However, we shall see that in the elaboration of the theory Lewis is forced to build more than extension into his interpretations.

In Convention, an interpretation is construed as an ordered pair consisting of a mood--indicative, imperative, etc.--and a truth condition--a set of possible worlds, those at which the truth condition holds in the sense appropriate to the mood of the sentence. The grammar assigns such a pair to each sentence of the language. But this basic idea needs to be complicated in various ways. Because natural languages may contain sentences that have indexical and anaphoric features, and that are thus dependent upon present context or prior conversational context, the grammar will have to be relativized so that it assigns interpretations to sentences upon possible occasions of utterance, an occasion of utterance being a possible world plus a spatio-temporal location therein. Furthermore, sentences may not be unambiguous; so a set of interpretations must be specified for each sentence. Thus a grammar turns out to be a function from pairs of sentences plus possible occasions of

utterance to finite, possibly empty, sets of interpretations (mood plus truth condition).

In "General Semantics"[6] Lewis fleshes out his conception of a grammar and introduces a more complex notion of an "interpretation." He describes what he calls categorially-based transformational grammars, modeled upon Ajdukiewicz categorial grammars, which are then to be supplemented with a transformational component. Again what he is after is "abstract semantic systems whereby symbols are associated with aspects of the world."[7]

What does such a system consist of? We begin with a small number of basic categories of expressions, for instance, the category S of sentences and the category N of names. (Which categories, and how many, are chosen as basic is a practical matter, so in this example I keep things to a bare minimum.) To these basic categories are added (infinitely many) functor categories of the form X/Y, whose expressions combine with an expression of category Y to form a compound expression of category X. Examples of these derived categories would be: S/N, the category of verbs (which combine with names to form sentences); N/N, the category of adjectives (which combine with names to form names); and (S/N)/(S/N), the category of adverbs (which combine with verbs to form verbs). Next a finite lexicon is

---

[6]"General Semantics," in Semantics of Natural Language, ed. Gilbert Harman and Donald Davidson (Dordrecht, Holland: Reidel, 1972), pp. 169-218.

[7]"General Semantics," 170.

specified, each member of which is assigned to a category, and finally the set of transformational rules is specified.

Given such a grammar, we are in a position to inquire about meanings of expressions in the language. For Lewis, the meaning of an expression remains that which determines its extension. Since the extension of an expression depends upon a variety of factors--circumstance of utterance, speaker, etc.--Lewis introduces a function whose domain will be that set of factors and whose range will be extensions of the appropriate kind: truth values for sentences,[8] things named for names, sets of things for common nouns. He labels this function an "intension," for, while it is not an intension in the standard sense, it is designed to do the work for which such intensions are intended. Specifically, a Lewisian intension is a function from indices to extensions, where an index is a finite sequence of coordinates, comprising a possible-world coordinate, contextual coordinates (time, place, speaker, indicated objects and previous discourse) and an assignment coordinate (assigning values to variables). The notion of an index may be thought of as an attempt to spell out what Lewis referred to in Convention as a possible occasion of utterance.

Only members of the basic categories have such intensions. Intensions for derived categories differ from

---

[8]Lewis, here concentrating on indicative sentences, relegates moods to a secondary role; they can be reintroduced at a later point in the development of the grammar.

those for basic categories: since members of derived categories may not have extensions, for them there can be no functions from indices to extensions. So Lewis defines their intensions compositionally as functions from intensions to intensions. The intension of a derived category will be a compositional function from the intensions of the items appearing to the right of the oblique stroke to the intension of the item appearing to its left. Syntactic compositionality is mirrored by semantic compositionality. (Where derived categories such as extensional adjectives do have extensions, intensions of the first sort may be defined. But for purposes of uniformity of treatment, Lewis disregards them.)

These intensions of Lewis's are evidently artificial constructs. The question then arises what relation they bear to the pre-theoretic idea of meanings. It is immediately clear that the two cannot be identified. Differences in intension capture only coarse differences in meaning. Tautologies differ in meaning but not in intension. Sameness of intension is adequate only for dealing with non-compound, lexical constituents. For fine differences, the intensions of the components of the expressions must be taken into consideration. Thus Lewis identifies meanings with "semantically interpreted phrase markers minus their terminal nodes [expressions]: finite ordered trees having at each node a category and an

appropriate intension."[9] The lexical part of the syntactic component of the total grammar will assign a lexical entry to each terminal node of the tree, producing a deep structure; and the transformational part will produce the surface structure of an expression of the language. The total grammar may now be defined as a relation between a set of expressions and their meanings.

A GRAMMAR

involves the assignment to

LEXICAL ITEMS

of

SYNTACTICAL CATEGORIES	INTENSIONS
followed by application of	followed by application of
SYNTACTICAL COMPOSITIONAL RULES	SEMANTICAL COMPOSITIONAL RULES
yielding	yielding

EXPRESSIONS <- - - - - - - - -> MEANINGS

Before going on to look at how grammars are assigned to actual languages, the question raised by Lewis's freewheeling use of intensions needs to be addressed: if his use of possible worlds does not place Lewis outside the Quinean extensionalist lineage, does his use of intensions lead to the same result? In his 1974 article

---

[9]"General Semantics," 182.

"'Tensions,"[10] Lewis offers a way of capturing the above intensional language in a purely extensional language. In the extensional transform of the above intensional language, we have S-names and N-names. The extension of an S-name is identical to the intension of the expression of category S (of the intensional language) that it names, that is to say, a function from indices to a truth value. The extension of an N-name is identical to the intension of the expression of category N (of the intensional language) that it names, that is to say, a function from indices to an entity of some sort. The extension of members of functor categories will be a compositional function of the extensions of its constituents. Thus anything expressible in the first language is expressible in the second, and extensionality is preserved by taking the intensions of the intensional language as the extensions of the second-order names of which the extensional language consists.

Up to this point we have been looking at grammars independently of their application to actual languages. Each such grammar determines what Lewis calls a possible language. However, he wishes to apply this approach to actual languages used by human populations. To do this he needs to identify the actual language used by a population with the appropriate possible language. This is the problem

---

[10]In Semantics and Philosophy, ed. Milton Munitz and Peter Unger (New York: New York University Press, 1974); reprinted in Philosophical Papers, vol. 1, 250-260.

of radical interpretation, or at least a part of it.[11] In Convention Lewis proposes a solution. Where Davidson appeals to a principle of charity that ascribes to native speakers our beliefs and makes of them truth-speakers, Lewis appeals to a convention of truthfulness: we ascribe to the native speakers the desire to be truthful, although success is not always guaranteed. By truthful Lewis means truthful and forthcoming, i.e., trying to avoid uttering falsehoods, but not simply by refusing to engage in conversation. Uniform truthfulness and corresponding trust on the part of hearers permit successful communication; they are its necessary support.

> If σ is a sentence of L which would be indicative in L on an occasion o of its utterance by x to an audience in P then x tries to make sure that he utters σ on o only if σ would be true in L on o.[12]

> L is an actual language of P if and only if there prevails in P a convention of truthfulness in L, sustained by an interest in communication.[13]

In "Radical Interpretation"[14] Lewis sets out more fully his view of the nature of the problem of radical interpretation; the constraints upon it; and the degree to

---

[11]In "The Structure and Content of Truth," Davidson now presents his theory of meaning in this two-step way: (a) construct truth theories; and (b) determine which truth theory applies to a given population of speakers (the task of radical interpretation, confirming the truth of T-sentences implied by a truth theory.

[12]Convention, 192.

[13]Convention, 194.

[14]Synthèse 27 (1974); reprinted in Philosophical Papers, vol. 1, 108-121.

which the problem is susceptible to a determinate solution. He casts the problem in terms of the interpretation of the beliefs and desires of an individual: given a physical theory P of an individual (call him Karl) and the world, to specify Karl's system of beliefs and desires in both his own terms A_k and in ours A_o and also to specify in our terms the truth conditions of his sentences M. (His formulation of the problem is broader than Davidson's: given A_k, solve for M and A_o.) However, the solution that Lewis proposes may be applied to the task of radical interpretation as we have been discussing it: how do we assign a grammar (syntax + semantics) to a language used by a population?

To the single principle--truthfulness--presented in Convention, Lewis adds five further constraints that are intended together to permit the solution of the problem of radical interpretation. The six are:

a. The Principle of Charity, which attributes a common inductive method and underlying system of basic intrinsic values to ourselves and to Karl (note that it does not simply attribute true beliefs to Karl, as Davidson would have us do, but is more nuanced);

b. The Rationalization Principle, which requires that belief and desire ascriptions be such as to make Karl a rational agent, one with good reasons for his behavior;

c. The Principle of Truthfulness, which makes of Karl

basically a truthful speaker;

d. The Principle of Generativity, requiring a finite, reasonably uniform and simple specification of the truth conditions of the sentences of Karl's language, as in the manner of Davidson or in the way outlined in "General Semantics";

e. The Manifestation Principle, requiring that Karl's beliefs and desires be manifest in his speech dispositions and behavior, i.e., that he not be deceptive; and

f. The Triangle Principle, requiring that our attribution of beliefs, desires and meanings be such as to make Karl's beliefs and desires be the same whichever language they are expressed in.

§3

It is time now to take stock of what Lewis's theory is designed to provide us in the way of answers to the problems that theories of meaning traditionally tackle. From the start Lewis departs from Quine in accepting semantic structure within modal contexts. 'Possibly . . .' and 'Necessarily . . .' are susceptible to semantic treatment in ways beyond those allowed by Quine in "Quantifiers and Propositional Attitudes" and "Three Grades of Modal Involvement,"[15] which simply try to isolate and condense

---

[15]In Proceedings of the XIth International Congress of Philosophy (Amsterdam: North-Holland, 1953); reprinted in The Ways of Paradox, 158-176.

the intractable modal element. As a direct result, Lewis subscribes to de re necessity and what Quine calls Aristotelian essentialism.

> An attribute that something shares with all its counterparts is an essential attribute of that thing, part of its essence. The whole of its essence is the intersection of its essential attributes, the attribute it shares with all and only its counterparts.[16]

As regards synonymy, two kinds of semantic value are assigned to each expression: (a) intension and (b) meaning, treated in Lewis's way as a tree-structure that captures the compositional genealogy of the intension assigned to an expression. Neither corresponds precisely to the pretheoretic notion of meaning. But judicious appeal to one or the other can serve to account for sameness and difference of meaning. When sameness of intension is too coarse to capture the distinction in meaning between, say, two analytic sentences, true in all possible worlds, difference of tree-structure may be invoked to draw the distinction. When sameness of tree-structure cuts too fine a distinction, as in the case of a sentence and its double negation, then sameness of intension suffices.

We also find propositions and analyticity rehabilitated. A proposition is taken to be a set of possible worlds, those in which the proposition holds. More generally, it is a property instantiated only by entire worlds, the property of being a world where that proposition

---

[16]"Counterpart Theory and Quantified Modal Logic," 35.

holds. Such propositions are coarse-grained, corresponding to intensions. The employment of propositions is to be restricted, however, to what is physicalistically determinable.

An analytic sentence is one that is true at every possible world, or, more precisely, true at every index. However, Lewis's restoration of analyticity does not succeed in distinguishing between necessary truths in general and those that are analytic, true specifically in virtue of meaning. The logical truths and the stereotypically analytic truths are all true in all possible worlds. And Lewis's application of the term 'analytic' to both types ignores the distinction. It may be argued that this is of little consequence if the theory can come through on the rest of what is promised. However, as we shall see later, analyticity is involved in further troubles for Lewis's theory.

What of indeterminacy? The convention of truthfulness is a limiting case. It does not in practice succeed in selecting one possible language as the actual language spoken by a particular population. Typically it can manage to select a cluster of languages. So, to the extent then that radical interpretation falls short of selecting one language, indeterminacy remains. The further constraints offered in "Radical Interpretation" still cannot eliminate indeterminacy. But what does it mean to say that there is indeterminacy? Lewis acknowledges a measure of

indeterminacy below the sentential level and also to the extent that no solution will fit all the constraints perfectly. However, he rejects the more radical idea of two incompatible solutions that would perfectly meet the constraints. Such a radical posture (Quine's) simply fails to recognize an alternative explanation, that one has failed to come up with all the constraints.

Inscrutability also remains. In order to tie a particular interpretation to a subsentential expression employed in a certain population, we need to mention a language and a grammar. What privileges a grammar? Different grammars interpret constituents differently. Conventions of truthfulness pertain only to whole sentences. If there is no method for determining the optimal grammar then inscrutability remains.

§4

It is clear that if Lewis's theory can deliver on all this, then it moves way beyond the limited offerings of Quine's theory. However, it is susceptible in all three areas: the account of modality upon which the semantics proper rests; the employment of possible worlds in the specification of grammars; and the account of radical interpretation, which is what allows the abstract grammars a hold on actual languages. I shall treat each area in turn.

For all the work that possible worlds are meant to do, Lewis introduces them as altogether innocuous. Possible worlds are ways this world might have turned out. They and

their inhabitants are just like this world and its inhabitants, the same kind of thing (and hence introduce no new kinds of entities that might violate parsimony strictures or the typical employment of the quantifiers). The only difference is that the world we are in actually exists, while other possible worlds exist but do not actually exist. They are not to be criticized for this shortcoming because 'actual' is an indexical referring to our world; to actually exist is to exist at our world, which, by definition, unactualized possible worlds do not do. They cannot be criticized for failing to do what they are not meant to do.

These claims are not as unexceptionable as Lewis would have us believe. To undermine Lewis's theory of meaning we need not prove that Lewis's account of modality is wrong; we need simply show that it is sufficiently problematic as to make it incapable of shedding any light on meaning. It is no use trying to explain the obscure by the more obscure.

A great deal turns upon Lewis's distinction between existence and actual existence. But are the words 'actual' in the expression 'actual existence' or 'actually' in the expression 'actually exists' doing any real work? They seem instead to be employed, not indexically as Lewis would have it, but rhetorically, for emphasis, in just the way that the word 'real' was employed in the expression 'real work' in the last sentence. If they are doing no work, then to say that possible worlds exist but do not actually exist sounds

like a flat-out contradiction. So some other way of differentiating this world and other possible worlds will have to be found. But if the existence-actuality distinction collapses into the possibility-existence distinction and existence returns to being the prerogative of the actual, then whatever new status is attributed to possibilia, it would be changing our use of the quantifier to have it range over things that do not exist and are merely possible.

Put another way, Lewis needs the word 'actual' to function indexically so that the earlier claim that there is no difference in kind between this world and other possible worlds can hold up. Otherwise, the difference must be a difference in description or in kind. If the former, then possibilia will no longer be concrete and independent, but abstract and linguistic or conceptual. If the latter, then possible worlds will not be the same kind of thing as the world we are in.

Now if there is no difference in kind between this world and other possible worlds, then what kind of thing is it that we are being asked to consider? It is important to know just what kind of thing a possible world is, for that is necessary to the indexical reading of 'actual'. If 'actual' is to function as an indexical, then it must serve to pick out one member from among a group of things of the same kind. When using an indexical we need to know what kind of thing is being indicated: is it that terrier, that

dog, that mammal, that animal, etc.? In the present case, we want to pick out the possible world (rather than: neighborhood, town, country, planet, etc.) we inhabit from among the possible worlds (rather than: possible neighborhoods, towns, countries, planets, etc.) we do not. Lewis's initial answer in Counterfactuals--"ways things might have been"--is unsatisfactory: a world is a thing and not a way, at least that is what Lewis wants to maintain.[17] Later in Counterfactuals, he writes:

> I can only ask [the questioner] to admit that he knows what sort of thing our actual world is, and then explain that other worlds are more things of that sort, differing not in kind but only in what goes on at them.[18]

But this is a virtual non-answer. We still need to know what kind of thing our world is. Let us essay an answer on the order of all those spatio-temporal things that are causally connected to me. This doesn't do much good either. Are causality and spatio-temporality essential to being a possible world? Might there not be possible worlds that exhibit neither property? Kant, for one, saw them as phenomenal, and not noumenal. Perhaps Lewis is mistaken about this world's exhibiting those two properties; perhaps he is overlooking those parts of this world that are not in

---

[17]Kripke's conception of possible worlds not as concrete entities but as abstract states, one and only one of which is realized, accords much better with this talk of "ways," which he, too, employs; see Preface to Naming and Necessity (Cambridge, Mass.: Harvard University Press, 1979), 15-20.

[18]Counterfactuals (Oxford: Basil Blackwell, 1973), 85.

the same space-time frame as he is and are not causally related to him. Where are the limits to be set on this world, let alone on other, merely possible worlds? The difficulty that arises in answering this question is that it is not one that standardly arises either for the scientist or for the layperson. The standard answer is that the world is everything that exists or everything there is, whence the absence of any need to specify its limits.[19]

As for Quine's demand for criteria of individuation and identity, Lewis responds to it with his notion of a counterpart, according to which the question when we are dealing with one possible individual in two possible worlds or different possible individuals that share a great many attributes does not arise, since no one thing is ever in two possible worlds. However, no sooner is the problem of trans-world identity dismissed than it is replaced by another problem: how do we determine whether and when an object is a counterpart of another? Lewis's response relies on comparative similarity, but determining degrees of overall similarity between individuals in different possible worlds is bound to be a difficult task and one whose products are going to be vague. Lewis's defense in Counterfactuals is that the notion is vague because what it is a notion of is vague, and indeed comparative similarity whether across worlds or within worlds is vague. So it

---

[19]I pass over the further obvious difficulty for Lewis that this standard answer denies his existence-actuality distinction.

should not be held against modal realism. "Truth conditions for counterfactuals are fixed only within rough limits . . . varying with every shift of context and interest,"[20] but no more so than comparative similarity. Now if the notorious vagueness of counterfactuals excuses the vagueness of Lewis's account of similarity relations, it does not excuse his putting this account to work in the underpinnings of a theory of meaning. The reason some semanticists stick with extensions is that senses and such are held to be notoriously difficult. Truth conditions seem much more tractable (such was Davidson's motivation: truth conditions in the form of T-sentences are pre-theoretically available). But in requiring that we apprehend truth conditions in all possible worlds, Lewis is asking us to eschew intensional entities in favor of extensions that he admits are notoriously problematic as well. It has not been made clear that one is any better off for doing so. To say that meaning is terribly fuzzy, and then to produce a fuzzy theory of meaning satisfies no one, neither those who wish to do away with meaning nor its staunch defenders.

The second area of criticism has to do with truth in possible worlds and analyticity. Lewis admits that while the definition of analyticity is clearcut, it is another question whether or not anything meets that definition. Analyticity, Lewis states in Convention, is "unsharp."[21]

---

[20]Counterfactuals, 92.

[21]P. 200.

Analytic sentences are no longer true so much in virtue of meaning as they are true in virtue of possible worlds. Sometimes we cannot tell whether a particular highly theoretic statement is true in a particular possible world. And this casts into doubt not simply analyticity but also our ability to determine whether a particular world is a possible world.

Let us assume that sentence s is a prime candidate for analytic status; it appears to be true in all possible worlds. But now someone notices a world that has not been considered; call it $w_{n'}$. $W_{n'}$ is identical to w_n except that s is true in w_n but false in $w_{n'}$. Is $w_{n'}$ not a possible world because s is false in it or should we revise our preliminary determination that s is true in all possible worlds?

The relation between truth in every possible world (and that of course is analyticity) and what determines whether a given world is possible appears circular. A sentence is analytic if and only if it is true in every possible world. A world is possible if and only if no negation of an analytic sentence is true in it. The first of these two statements Lewis accepts as definitional of analyticity. The second he must reject as not definitional of possible-worldhood on pain of circularity; on Lewis's account, possible worlds are not simply models of the analytic sentences. And the circularity is threatening, because we are not given any other way to determine whether a possible world is possible, for possible worlds are primitive. It is

the same circle Quine identified in "Two Dogmas."

We cannot simply assert that $\underline{s}$ is analytic and true and therefore $w_{n'}$ must not be a possible world, because that is reversing the relationship between analyticity and possible-worldhood. We have no independent grasp of analyticity with which to fix the possible worlds. It is the possible worlds that must fix the analytic sentences.

Finally the trouble with Lewis's proposed solution to the problem of radical interpretation is that it is supported by little in the way of argument. Lewis we have seen offers six principles, some of which are more directly connected to meaning than others. But it is a holistic enterprise that ascribes meaning and belief simultaneously. What is the status of these six principles? Earlier Lewis argued for one of them as a convention, that is to say as a regularity in the behavior of the members of a population conditioned by universal conformity and expectation of conformity, and universal knowledge of this situation. This conformity is sustained by a mutual interest on the part of the group. A linguistic convention, then, is a regular form of speech behavior sustained in a group by a mutual interest in communication. Such is the convention of truthfulness. We can accept generativity as a result of human finite learning capacity. But what of the other principles advanced in "Radical Interpretation?"

Lewis states that they are:

> the fundamental principles of our general theory of persons. They tell us how beliefs and desires

> and meanings are normally related to one another, to behavioral output, and to sensory input.[22]

They implicitly define a person's belief-desire-meaning system: nothing is a belief-desire-meaning system unless it conforms to the principles; and humans do in fact have such systems. The principles are derived from common sense. As such, they have "a status akin to analyticity,"[23] although Lewis pulls back from committing himself fully to this claim. In fact Lewis diminishes his claim considerably. The six principles are not intended to eradicate indeterminacy. They select a set of grammars, not a single one. And they may not be enough to do the job; supplementary principles may be needed. What kind of case is this for the success of these principles in providing a solution to the problem of radical interpretation? At best, they narrow the scope of the indeterminacy. At worst, they remain the platitudes that Lewis expects from common sense. Why should common sense be thought capable of selecting, by a series of successive applications of these principles, a set of grammars for the actual language used by a population? The distance between common sense and the abstract grammars Lewis has developed is large enough to warrant more explanation of their connection than Lewis's suggestions about the order of application of the principles.

---

[22]P. 111.

[23]"Radical Interpretation," 112.

So it is uncertain how much indeterminacy is reduced. And even before that point is reached, the theory has to overcome the difficulties concerning the ontological status of unactualized possibles, the individuation of counterparts and the specification of possible worlds. Curiously, these difficulties are no different from those that Quine flagged in the first place.

CHAPTER V

PUTNAM AND STEREOTYPES

Quine may be viewed as challenging the semantic theorist to isolate the independent contributions of semantic information and empirical information to linguistic behavior. Davidson, as we have seen, takes up this challenge directly, but fails to meet it. Putnam, in contrast, aims to circumvent it, for he takes Quine's problem of separating meaning and belief to be a false one, merely the latest avatar of a series of dichotomies running throughout the history of the theory of meaning: intension and extension, sense and reference, analytic and synthetic, dictionary and encyclopedia, meaning change and theory change.

What Putnam sets out to do is rid the theory of meaning, once and for all, of this recurrent dualism. Establishing these distinctions is not the goal of semantics, but an obstacle to it. That they fail to hold up under scrutiny should be cause for neither surprise nor despair. Instead their deficiency is the central fact that needs to be recognized before a theory of meaning can be built.

In particular, Putnam seeks to eradicate the dichotomy between the private and public aspects of language, or, as

he puts it, what's in the head and what's outside the head. Adherence to this distinction prevents the recognition that the use of language is a social and not just a psychological art. Even though intensionalists, following Frege, are critics of psychologism, they continue to assume that grasping the meaning of an expression is an individual psychological act, that knowing the meaning of an expression is simply a matter of being in a certain solipsistic mental state. That is to say, while intensions are not psychological entities, our relation to them is described by a psychology, descended from Descartes, that, in the specification of mental states, restricts itself to proximal inputs and outputs without regard for the external source of those inputs and outputs.

What's more, Putnam maintains, it is from his intensionalist foes that Quine has inherited this Cartesian epistemological error of specifying mental states narrowly. Regardless of the differing ontological status that they attribute to meanings, they all fall victim to the same solipsistic assumption. There is, Putnam contends, no reason for extensionalists to accept its constraints. Rejecting this error in favor of a robust realism that recognizes the role of the external world--things and speakers--in the meanings that our words take on is the way to avoid Quine's scepticism about meaning and to construct a robust semantics.

In light of Putnam's antipathy toward dichotomies, it

may well strike his readers as odd that the one dichotomy he does not do away with altogether is the analytic-synthetic distinction, often taken to be the foundation of the other distinctions. Rather than champion or simply tolerate the distinction, however, Putnam strips it of all significance. His treatment of the analytic-synthetic distinction will be the subject of the first section of this chapter. It is of twofold interest. First, in it we find the beginnings of the theory of meaning that Putnam later advances. Second, an examination of Putnam's account of analyticity and the problems with it will prepare us for similar problems with his theory of meaning: what presents itself initially as a new and interesting position turns out to have less to it than meets the eye.

The second section of this chapter, relying principally on "Is Semantics Possible?"[1] and "The Meaning of 'Meaning',"[2] will lay out the theory. In the third section, I shall enumerate the specific points on which Putnam's theory distinguishes itself from Quine's. Then in the fourth, I shall argue that these departures do not constitute a viable theory of meaning that moves beyond

---

[1]In Languages, Belief and Metaphysics, vol. 1, Contemporary Philosophic Thought, ed. H. Kiefer and M. Munitz (New York: SUNY Press, 1970); reprinted in Mind, Language and Reality, vol. 2, Philosophical Papers (Cambridge: Cambridge University Press, 1975), 139-152.

[2]In Language, Mind and Knowledge, vol. 7, Minnesota Studies in the Philosophy of Science, ed. Keith Gunderson (Minneapolis: University of Minnesota Press, 1975); reprinted in Mind, Language and Reality, 215-271.

Quine's: of the two central elements of Putnam's theory, one, the causal theory of reference, is in need of an explanation of the semantic content of intentional states, while the second, the theory of stereotypes, in addition to having internal problems, represents no real advance upon Quine's position, since espousing stereotypes would not commit Quine to anything not already implicit in that position. Moreover, it is possible to raise the question whether Putnam presents us with a theory of meaning that is a theory.

§1

In his 1962 article "The Analytic and the Synthetic,"[3] Putnam takes a position that is in disagreement with both the detractors and the defenders of the analytic-synthetic distinction: he maintains, against the former, that there do exist analytic truths and, against the latter, that this fact is trivial, since nothing of philosophical consequence follows from their existence.

Putnam acknowledges that those sentences typically presented as obvious examples of analyticity, e.g., 'All bachelors are unmarried', are indeed analytic. But what makes these sentences analytic is so particular that analyticity is but an odd quirk exhibited by a limited

---

[3]In Scientific Explanation, Space, and Time, vol. 3, Minnesota Studies in the Philosophy of Science, ed. Herbert Feigl and Grover Maxwell (Minneapolis: University of Minnesota Press, 1962); reprinted in Mind, Language and Reality, 33-69.

number of sentences, and nothing upon which to found a whole theory of meaning. The analyticity of 'All bachelors are unmarried' depends upon the particular nature of the subject expression, 'bachelor'. A person is a bachelor if and only if that person is a man who has never been married. This biconditional serves as a criterion, the sole criterion, that determines the application of the term 'bachelor', for which reason Putnam labels such terms "one-criterion" words. Their application depends upon a single set of necessary and sufficient conditions, in virtue of which, Putnam maintains, they support analytic sentences. The error committed by defenders of the analytic-synthetic distinction is to let themselves be persuaded by the tidy semantic treatment of one-criterion words that all common nouns can be handled in the same way and thus support analytic sentences.

And this overenthusiasm about the powers of analyticity is an error, in Putnam's estimation, of far graver consequence than Quine's repudiation of it. 'All bachelors are unmarried' may be treated as true <u>a priori</u> in virtue of meaning, or the rules of language, or what have you, because there is a single criterion that determines its application. As such, it is immune to revision short of change of meaning or change of linguistic and logical rules. In contrast to such analytic sentences are those sentences refutable by isolated experiments, e.g., 'My blue pen is hidden under the papers on my desk'. One cannot, however, take all sentences that are not refutable by isolated experiments to be

analytic; their unrevisability may be due far more to their importance in one's conceptual scheme than to logico-linguistic convention; and it is a mistake to conflate systematic import and conventionality. The continuum of sentences does not divide neatly, barring some fuzziness where the two meet, into the analytic and the synthetic. Rather there are the analytic sentences at one end, the synthetic at the other and then, in between, all the really interesting sentences, those that present a challenge to the semantic theorist.

So Putnam holds that while Quine is wrong about the existence of the analytic-synthetic distinction, he is right in warning against its overuse and abuse. Trying to attribute analyticity to a far broader range of sentences doesn't work; not all language use is governed by criteria. While the distinction exists, it had best be ignored. Quine's essential point, that no statements are immune from revision before the tribunal of experience, holds good, Putnam contends, for all but that small subset of trivial analytic statements.

Putnam should not, however, place a limit on the set of revisable statements. For on his account of analytic sentences, they are not immune from revision. The biconditionals that support them are themselves contingent. And if the truth of Putnam's analytic sentences is dependent upon that of contingent sentences, then they are themselves contingent, and hence open to revision. Moreover, analytic

sentences are supposed to be necessary. And Putnam's analytic sentences are not necessary, so not really analytic. There is something missing in Putnam's account: it fails to make a case for the analyticity of his analytic sentences.

So what initially presents itself as a new and interesting position in the debate about the analytic-synthetic distinction collapses back onto Quine's position that there is no such thing as analyticity. And this is a precursor of things to come: Putnam's later theory of meaning, which appears, at least on its face, to constitute a new and interesting approach, collapses back onto Quine's theory: what substance there is to the new theory is borrowed from Quine's.

§2

Since, as Putnam states, his analytic sentences are intended to do no work in the theory of meaning anyway, the theory does not stand or fall on this matter. So let us turn to a direct examination of it. Putnam sets out by examining natural kind terms,

> general names associated with . . . classes of things that we regard as of explanatory importance; classes whose normal characteristics are "held together" or even explained by deep-lying mechanisms[, e.g.,] gold, lemon, tiger, acid.[4]

Indeed, a significant part of his thinking about meaning is devoted to natural kind terms; from his critique of the

---

[4]"Is Semantics Possible?" 139.

analytic-synthetic distinction on through his Twin Earth examples, he keeps returning to them. To get a handle on them he uses one-criterion terms.

Now one-criterion terms such as 'bachelor' do not occur essentially in any empirical generalizations; they are of no explanatory import and thus may be treated as mere dispensable abbreviation. Natural kind terms, unlike 'bachelor', are of explanatory import and nowhere near as clear-cut in their usage. Their semantics, Putnam argues, cannot be assimilated to that of one-criterion terms. Their application does not turn upon single criteria; rather they are "cluster" terms. Their use--and this is the source of their role in explanation--is determined by their figuring in a cluster of generalizations, none of which is necessary or sufficient, that encompass both semantic information and embedded collateral information of an empirical nature. Putnam initially talks of "law-cluster" terms, but later withdraws this reference to laws; for speakers of a language can hardly be held responsible for a knowledge of empirical laws. But he retains the idea that the use of natural kind terms is determined by a cluster of law-like generalizations, in which we can see a fore-runner of his later notion of stereotype.

Natural kind terms serve in Putnam's thinking a purpose that is not simply negative--warning semantic theorists away from a too facile reliance upon analyticity--but also a more important, positive one. They present a prime case of the

commingling of linguistic and empirical information, the very thing that Quine saw as a stumbling block for semantics and that Davidson sought to overcome. Putnam, however, sees this not as an impediment but as an insight afforded us by natural kind terms and essential to a proper understanding of meaning. The traditional goal of winnowing real "meaning" out of this amalgam is a false one. In practice, when you learn the meaning of a new word, what you are after is the information that will permit you to use the word correctly. That such information fails to divide up neatly into dictionary and encyclopedia entries, or analytic and synthetic, is just too bad. Where is the scientific merit in sticking to some theoretical abstraction rather than to the data? The point that Putnam wishes to drive home is that the mixture of linguistic and empirical information is the best candidate for the object of study of the meaning theorist.

It comes as no surprise then that natural kind terms figure prominently in Putnam's 1970 article "Is Semantics Possible?" which presents in its developmental stages the theory of meaning that will be fully laid out in his 1975 "The Meaning of 'Meaning'." In the earlier article, Putnam reiterates his diagnosis of the error of traditional theories of meaning down through Carnap and Katz; they push analyticity beyond its proper limits and wind up misrepresenting the semantics of natural kind terms:

> the traditional theory has taken an account which is correct for the 'one-criterion' concepts (i.e.

> for such concepts as 'bachelor' and 'vixen'), and made it a general account of the meaning of general names. A theory which correctly describes the behavior of perhaps three hundred words has been asserted to correctly describe the behavior of the tens of thousands of general names.[5]

But if traditional theories of meaning are mistaken, what alternative is there? Is semantics impossible? Putnam refuses to subscribe to Quine's pessimism. But the question is, where to begin? On what can a theory of meaning be built? Putnam takes as his point of departure the practical fact, alluded to above, that people can be taught the use of new words, and quite easily at that. "[I]n this simple phenomenon lies the problem, and hence the raison d'être, of 'semantic theory'."[6]

> The fact that one can acquire the use of an indefinite number of new words, and on the basis of simple "statements of what they mean," is an amazing fact: it is the fact, I repeat, on which semantic theory rests.[7]

A caveat needs to be made about these remarks: any temptation to draw a parallel between Putnam and Davidson here should be resisted. It is true that Putnam talks about a kind of learnability and some kind of indefinitely large linguistic capacity. But we ought not make too much out of this. There is no Chomskian or Tarskian subtext here. Putnam is talking not about sentences, but about words. It is, as he himself insists, a simple, even naive observation.

---

[5]"Is Semantics Possible?" 141.

[6]"Is Semantics Possible?" 147.

[7]"Is Semantics Possible?" 149.

But that is what Putnam wants in the way of a theory of meaning. He is not bent upon abstracting properly linguistic information out of the assorted information we absorb when we learn the use of a new word. He wants to clear the decks and employ common sense.

The approach that Putnam recommends immediately rules out Quine's proposal to construct a semantics on the basis of stimulus meaning. Stimulus meaning is not the kind of thing that is conveyed when people teach or learn new words; it is too complex a notion to be transmitted as straightforwardly as meaning seems to be conveyed in everyday life.

"Is Semantics Possible?" presents us with two new notions: core facts and stereotypes. About the former Putnam advances two empirical claims:

> there are, in connection with almost any word . . . , certain core facts such that (1) one cannot convey the normal use of the word . . . without conveying those core facts, and (2) in the case of many words and many speakers, conveying those core facts is sufficient to convey at least an approximation to the normal use.[8]

These core facts constitute the meaning of the word; they are what you supply when you give the meaning of a word. They do not carry a set of necessary and sufficient conditions for identifying the extension of the word; they do not constitute intensions. There is no attempt to separate out properly linguistic and properly empirical facts. The task of the semantic theorist will be to

---

[8]"Is Semantics Possible?" 148.

describe the various kinds of core facts associated with each kind of expression. In the case of one-criterion words the core facts are simply the criterion itself. In the case of natural kind words, the core facts are the stereotype associated with the word, an oversimplified theory describing a normal member of the kind, plus the extension.

To look more closely at the notion of stereotype, we should turn to "The Meaning of 'Meaning'." In that article, the notion of core fact is discarded and that of stereotype takes center stage. It is in "The Meaning of 'Meaning'" that Putnam provides arguments for the theory suggested in "Is Semantics Possible?" What we need to take from the earlier article is the idea that Putnam sees a theory of meaning as tied to the task of explaining how communication and language understanding are possible. It is these elements that inform the theory of meaning. We are already being prepared for the idea that language is a public matter, a social phenomenon and not a private one. These are insights that Putnam will seize upon in the later article.

In "The Meaning of 'Meaning'," Putnam exposes the shortcomings of a solipsistic psychological approach to language learning and understanding. Recall that Putnam ascribes this mistaken view not only to those who espouse a frankly mentalist account of meaning but also to ostensibly anti-psychologistic philosophers such as Frege and Carnap, for whom knowing the meaning of an expression is to be in a

(narrowly-specified) psychological state. Such an approach makes it impossible to construct a semantic theory in which meaning determines reference, for no solipsistic mental state determines reference.[9] This is what the Twin Earth cases are intended to demonstrate.

The first Twin Earth case runs as follows: Suppose I have a Doppelgänger, molecule for molecule identical to me, on another planet identical to the Earth in every way except for the fact that on Twin Earth the stuff that looks and tastes and behaves just like water and is referred to by the word 'water' has the chemical composition XYZ, and not H_2O. When my Doppelgänger and I utter the word 'water', his mental state and my mental state are, by solipsistic hypothesis, identical; yet while I, in using the word, refer to water, namely, H_2O, my Doppelgänger on Twin Earth refers to XYZ. If sameness of meaning implies sameness of reference, then my utterance of the word 'water' and my Doppelgänger's utterance of it cannot have the same meaning even though our mental states are identical. Hence, knowing the meaning of a word cannot be a function purely of being in a certain solipsistic psychological state.

Putnam's explanation of this result appeals to "an unnoticed indexical component: 'water' is stuff that bears a

---

[9]It is Putnam's contention that Quine, while properly recognizing this consequence, concludes from it that we should be sceptics about meaning, whereas we should simply be sceptics about the psychology being employed.

certain similarity relation to the water around here."[10] Natural kind terms such as 'water' function in part the way demonstratives and proper names do. When I give the meaning of a natural kind term by means of a description, what I am doing is rather like giving an ostensive definition. When I say, "This is an X," my accompanying act of ostension serves to pick out some typical or standard example of an X. Of course, for the definition to succeed, I need the right setting. Also a little prompting may be in order: I need to provide information about what kind of thing an X is and whether 'X' is a verb, noun, adjective, etc., so that the person learning the word has some idea what ballpark we're in. In short, the referent of the expression is integral to its meaning.

When I give the meaning via a description, I replace the pointing and the setting with "a standardized description of features of the kind that are typical, or 'normal', or at any rate stereotypical."[11] The description serves as a stereotype that picks out "the water around here." It may succeed in this task despite being incomplete and possibly inaccurate. It is here, at the point where stereotypes begin to slip their factual moorings, that talk of core facts becomes strained. Knowing the meaning of the word 'water' requires knowing a certain standard minimum of information, the stereotype associated with it (not:

---

[10]"The Meaning of 'Meaning'," 234.

[11]"The Meaning of 'Meaning'," 230.

analytically predicated of it). It does not require knowing any set of necessary and sufficient conditions for something's being water. The set of necessary and sufficient conditions that does determine the extension of the term is not part of what the speaker must know in order correctly to be said to grasp the meaning of the word.

The second Twin Earth case is designed to show that this responsibility is shared, a state of affairs that Putnam labels the linguistic division of labor: Suppose that on Twin Earth those kitchen products, implements, etc. that on Earth are made of aluminum are instead made of molybdenum, even though Twin Earth's inhabitants call those objects "aluminum foil," "aluminum pots and pans," etc. Suppose further that my Doppelgänger has as vague an idea about the nature of aluminum and molybdenum, and the difference between them, as I do. It is nevertheless the case that when the two of us use the word 'aluminum', I refer to aluminum, while my Doppelgänger refers to molybdenum. Once again it appears that knowing the meaning of a word cannot be identified with being in a certain solipsistically-determined mental state. Putnam concludes that the burden of determining the extension of the word 'aluminum' is not shouldered by each speaker in isolation, but by the linguistic community as a whole, and, in this particular case, by its members who are engaged in the empirical study of metals.

The key then to the richer theory of meaning that

Putnam envisages is breaking the link between solipsistic mental state and meaning. Proclaiming one's anti-psychologism is insufficient. The semantic theorist, says Putnam, needs to adopt a realist position and recognize the role of the world and of the speech community in the determination of extension. His positive theory is that the meaning of a word should be described by:

> a finite sequence, or 'vector', whose components should certainly include the following . . . : (1) the syntactic markers that apply to the word, e.g. 'noun'; (2) the semantic markers that apply to the word, e.g. 'animal', 'period of time'; (3) a description of the additional features of the stereotype, if any; (4) a description of the extension.[12]

The first three components, which together constitute the stereotype, are what the speaker knows, what must be conveyed in giving the meaning of the word, what the speaker is in turn responsible for, the speaker's competence. The fourth component, extension, falls outside the range of the speaker's competence.

Now if the determination of reference is not within the speaker's competence, if speakers are not to be held responsible for fixing the extension of every expression they use, then on what grounds may it be said that the speaker succeeds in referring? How do we ascertain, for instance, that the word 'snow' refers to snow? Putnam's proposal to adapt the causal account of the reference relation situates the answer to these questions in the

---

[12]"The Meaning of 'Meaning'," 269.

genealogy/history of the word and the causal role that snow plays in that genealogy/history.

In assimilating the treatment of the extension of common names to that of proper names, Putnam reverses the maneuver of description theorists. Description theorists start from the position that common names are synonymous with a set of necessary and sufficient conditions that determine the reference of the name and then go on to treat proper names as disguised descriptions that refer to whatever the description is true of. The description theory runs into trouble when it turns out that a proper name may succeed in referring despite the fact that the description associated with it may fail to specify any object uniquely, may be satisfied by a different object or may be satisfied by no object at all. Similar problems recur with common names. Only one-criterion words refer to whatever the set of necessary and sufficient conditions associated with them is true of; and they are a special, non-generalizable case. The error of the description theorists, like that of the defenders of analyticity, is over-reliance on one-criterion words.

The causal theory can be presented from two points of view: that of a speaker's use of a word, and that of the word itself in the language. In either case, the reference of natural kind terms is modeled on that of proper names. When a speaker uses the word 'snow', he or she succeeds in referring to snow only if snow itself has played an

appropriate causal role in the speaker's acquisition of the word. It is the task of the linguist to trace that causal connection. If the speaker learnt the word in the presence of snow, then given the right kind of stage-setting (the speaker's attention is drawn to the snow) and prompting ("That stuff over there is snow"), the task is straightforward, for a direct causal connection has been established between the word and its reference, snow itself. And the speaker's use of the word succeeds in referring to snow.

If, however, the speaker has never seen snow, no direct causal connection is possible. Yet causal connection there must be, so the causal story becomes more complicated. The burden of successful reference is now shared by the person who, by saying something along the lines of "Snow is soft, white, crystalline flakes that fall from the clouds instead of rain in places where the temperature is below freezing," teaches the speaker the word 'snow'. Provided that nothing goes awry in the teaching, the speaker inherits the teacher's reference. Now if the teacher learnt the expression 'snow' in the presence of snow, an indirect causal chain linking the original speaker's use of the expression and snow itself is established. If the reference of the teacher's use of the word is also inherited, then the linguist must trace the causal chain to its source, an event at which the word 'snow' is introduced and snow is present. (Here we see Putnam's linguistic division of labor at work.)

The exact nature of this causal chain need not be known to the speaker, the speaker may forget the circumstances of learning the word, the speaker need not know the exact nature of snow, and yet the word 'snow' in the speaker's idiolect succeeds in referring to snow. In fact, the exact extension of the word need be known to no one, and in any case, its determination is a task not for the speaker, and not for linguists, but for the empirical scientists, meteorologists, who study the nature of snow.

The causal theory may also be viewed from a second perspective, that of the reference of the word in the language. Here the focus is on the introduction of the word into the language and not on an individual speaker's introduction to the word. The genealogical/historical method is once again employed. The sociolinguist, exploiting the knowledge gained by empirical scientists, traces the causal history of the expression back to its introduction into the language, Kripke's "initial baptism." (Obviously, in practice the linguist will not proceed in this fashion. But we are here examining the theoretical basis, and not the practical application, of Putnam's proposal.)

What is thus established is "a chain of linked referring-uses, a history of referential use going back in time,"[13] which the successful user need not be able to

---

[13]"Possibility and Necessity," in Realism and Reason, vol. 3, Philosophical Papers (Cambridge: Cambridge University Press, 1983), 57.

specify. The successful user need not be able to specify what constitutes membership in the same kind; extension "is not fixed by a set of 'criteria' laid down in advance, but is, in part, fixed by the world."[14] It is only afterwards that the appropriate experts perform their task of finding out just what that extension is.

§3

We are now in a position to assess the advances that Putnam's theory can claim over Quine's minimalist semantics. What do we want from a theory of meaning that Putnam's theory may be able to supply? The list of areas where we may seek advances looks impressive: truths in virtue of meaning; synonymy; radical translation; theory comparison; communication and interpretation; necessity; and referential scrutability. What remains of Quine's position is a commitment to holism, an acceptance of indeterminacy, albeit more limited, and the revisability of all statements.

The first of these areas we have already considered. On Putnam's account, there do exist truths in virtue of meaning. Had Putnam's account of the analytic-synthetic distinction survived, we would have been assured of this, with the truth of 'All bachelors are unmarried' being underwritten by the meaning of 'bachelor'. But in the absence of analytic truths there will be other truths that hold in virtue of meaning, namely those truths that hold in

---

[14]"Reference and Truth," in Realism and Reason, 71.

virtue of stereotype; such truths are not analytic and comport no necessity.

But the more important departures are only now evident. Putnam's account of meaning, with several elements going to constitute the meaning of a term, permits sense to be made of synonymy, commensurability and scrutability. Matching stereotype and reference provides grounds for the attribution of synonymy to a pair of expressions. Even if synonymy has to be relativized and turned into a matter of degree, there is a basis upon which to judge how close, or how far apart, in meaning two expressions are. And the lot of the radical translator is ameliorated. Matching stereotypes across languages provides a further constraint upon translation manuals, and thus diminishes the extent of indeterminacy.

The notion of stereotype applies not only to objects but also to properties. The description used to pick out a property is not analytically predicated of it; it is not to be viewed as constitutive of the property; it does not provide an operational definition of the property. Thus rival theories having incompatible descriptions associated with an expression for a physical property may yet be disagreeing about the same thing. Incompatible stereotypes do not preclude identity of reference. Where reference remains the same despite change of stereotype, there is common ground for rival theories to meet.

This is not the only place where Putnam's realism comes

into play. Problems of communication become more tractable, too, once it is recognized that meanings reside "outside the head." In fact communication and language understanding are the starting point from which Putnam has set out to construct his theory. On Putnam's account, interpreting the speech of others no longer requires determining the extension of the expressions they use. Nor does it require ferreting out psychological facts via behavioral interpretation. The grasping of stereotypes or "core facts" is sufficient, and these are publicly available.

Furthermore, entailed by the causal theory of reference is the need for de re necessity. Whatever the experts discover about the nature of natural kinds, rigidly designated by natural kind terms, is necessarily true of the natural kind, part of its "Aristotelian" essence, not of any description-dependent, nominal essence. Reference must be scrutable for the causal theory to get off the ground; the experts discover essences of natural kinds, not of undetached natural kind parts, natural kind time-slices, etc.

But all these advances are contingent upon the viability of Putnam's theory and its two pillars, his theory of reference and his theory of stereotypes. I turn now to that question.

§4

Let us consider first the account of reference. From both the perspective of speaker's use and that of the word

in the language, the problem with Putnam's theory of reference remains the same. The theory of reference falls within the domain of sociolinguistics. But in reconstructing an expression's genealogy the sociolinguist is not thereby explaining how reference works, but rather takes the reference relation for granted, simply seeking to assign the correct referent to the term.

Explaining how the initial baptism succeeds in establishing the reference relation, and just what this reference relation is, lies beyond the scope of sociolinguistics. For the success of the initial baptism depends, at least in part, upon the intentional state of the baptizer. In the course of the initial baptism, the expression receives its extension from the intentional state of the baptizer. Thus the reference of expressions in the language is inherited from speaker's (in this case, baptizer's) reference. Likewise, in the case of an individual speaker's acquisition of an expression, intentional state of both the speaker and the teacher are central to the successful transmission of the expression and its reference. But Putnam offers no theory of speaker's reference, no account of the semantic content of the intentional states that form the links in the causal chains. Without such an account,[15] the causal theory belongs in the realm of diachronic linguistics, like a history of kings of

---

[15]It is just this kind of account that Harman seeks to develop.

England and perhaps the circumstances of their succession, but devoid of any account of what monarchy is. What is explained is how, through the vagaries of history, Charles, rather than Joe, came to be king of England; or how, through the vagaries of history, 'Moses' came to refer to Moses, and not Aaron.

Let us turn now to the theory of stereotypes. Quine, we recall, has argued that no principled distinction can be drawn between truths that hold in virtue of meaning, on the one hand, and empirical truths, on the other: since no (non-logical) truths are immune to revision, all (non-logical) truths are on an equal footing. What Putnam has sought to do with the notion of a stereotype is to privilege a certain set of revisable truths and label them as semantic in virtue of the central role they play in communication and in teaching and learning how to use an expression. It is true that some truths are more privileged than others, that some will be revised more readily than others, and that some are highly resistant to revision. But these remarks are of only practical or perhaps psychological interest. What is it in the nature of these truths, not simply in our behavior toward them, that justifies privileging them in this way? How do we distinguish between those truths that belong to the stereotype of a word and those that do not?

Psychological evidence in support of stereotype-based

theories of meaning was advanced in 1973 by Eleanor Rosch.[16] Rosch's experiments showed that judgments about whether something counts as, say, a vegetable or a sport are typically graded. Football is consistently judged to be more of a sport than archery. These results were interpreted as providing support for the stereotype theory of meaning; whether something is judged to be a sport is a matter of the degree to which it resembles stereotypic sports, such as football and baseball.

A 1983 article by Armstrong, Gleitman and Gleitman, "What Some Concepts Might Not Be,"[17] calls into question this interpretation of Rosch's experimental findings. Armstrong et al. reproduce Rosch's results not only for those concepts that intuitively lend themselves to the stereotype view but also for those concepts that are well-defined, what Putnam has called "one-criterion" concepts, such as "odd number." Graded judgments are standard for both stereotypic and well-defined categories. The upshot is that graded judgments do not provide evidence that certain concepts fit the stereotype view while others do not, because graded judgments are the norm across the board.

The speculative conclusion at which Armstrong et al. arrive is that graded judgments are evidence for the

---

[16]See "On the Internal Structure of Perceptual and Semantic Categories," in Cognitive Development and the Acquisition of Language, ed. Timothy E. Moore (New York: Academic Press, 1973), 111-144.

[17]Cognition 13 (1983): 263-308.

existence of a heuristic identification function associated with a concept. That this identification function is the meaning of the word does not follow. Moreover, stereotype theories suffer from some of the problems that traditional, intensionalist theories have been criticized for. Some theory of the features that go to make up a stereotype needs to be provided; and here the question of primitive features arises, as much a problem for the stereotype theory as for the intensionalist theory.

Besides the problems that are shared, there are the problems specific to the stereotype view. Some metric for greater and lesser conformity to the stereotype needs to be provided. But no standard metric would seem to encompass resemblance to football (stereotypic sport) and carrots (stereotypic vegetable). That further research might yield some common set of parameters is implausible in the extreme.

Thus it is apparent that Putnam, in trying to avoid the problems with the analytic-synthetic distinction, has managed simply to reproduce them. This time the problems arise at the border between stereotype and non-stereotype, rather than at the border between analytic and synthetic. Whereas the first problem had been making the case for the analyticity of the analytic statements, now the problem is making the case for the stereotypicality of the empirical information to be included in the stereotype.

The concept of stereotype is also not strong enough to function as a standard of synonymy. Thus one of the claims

that may be made for Putnam's theory falls by the wayside. Stereotype can provide only degree of synonymy; it admits only of similarity, not of identity. And similarity is inadequate to the task of translation. Two terms may be perfectly good translations of each other despite the fact that they have different stereotypes. Consider for example the stereotypes associated with the English word 'priest' and its translation in pre-colonial X-land. The Catholic priest and the Xian priest both perform ritual functions, but dress, demeanor and lifestyle (simple cassock, meek and celibate vs. ornamented animal skin, fierce and polygamous) are polar opposites. And we can imagine the converse circumstance occurring--same stereotype, but poor translation pair.

Only in the extreme case of one-criterion words, artificial-kind words, where the stereotype is identified with the sole criterion, do we appear to get more than degree of synonymy. But similar misgivings arise in this case, too. How do you know you've got a one-criterion word on your hands? Again it is a matter of degree; there is no radical cut. So the same old problem recurs yet again. Some division is needed.

Troubles with one-criterion words in turn spell even more trouble for analyticity. On Putnam's account, the existence of analytic truths turns upon that of one-criterion words. The analytic truths are not unrevisable; they are defined by Putnam as those that can be obtained by

substituting synonyms for synonyms. One-criterion words do allow of synonymy. Or do they? If there is no precise way to determine whether an expression is a one-criterion word, then even that small number of obviously analytic truths that Putnam allows are less securely separated from the non-analytic. Putnam's theory has to allow more and less synonymy; the one exception is in the case of the analytic truths, where precision is provided by one-criterion words; but if one criterion words admit of more and less themselves, then the analytic truths will be more or less analytic. So even that ever-shrinking domain of truly privileged truths is insecure. Even the analytic truths do not provide us with out-and-out synonymy. And if synonymy is on the junk heap, then the aid that stereotypes are supposed to bring to the radical translator is illusory.

But Putnam can respond that his point all along has been the error of always seeking sharp distinctions. The notion of stereotypes is supposed to replace these unobtainable sharp distinctions. They are a mechanism by which some order and hierarchization can be restored to the arena of meaning. The stereotype theory is not meant to do the work that the old theory of meaning was intended to do. It simply provides a way of organizing truths into those that are more central and those that are less central.

But if that is all there is to Putnam's theory of stereotypes, then there is nothing in it that need be objected to by Quine, who clearly recognizes that some

truths are more equal than others. What label one chooses to employ is neither here nor there. This is where we started with Putnam. He appears to be making a major claim in reviving the analytic-synthetic distinction; but it turns out that his concept of analyticity is so weakened that there is nothing in it to which Quine need bother to object, barring perhaps the employment of a term that he may find repugnant. If the causal theory is merely the history of the continuities and discontinuities in linguistic usage, then it may be good history. If synonymy is not salvaged, then Quine again need not object. If a term's meaningfulness amounts to its having associated with it a vaguely delimited set of empirical beliefs, then again there is no reason for Quine to object.

A further problem with Putnam's theory is that it just may not be a theory of meaning at all. Putnam does not go beyond the organization of observations about how people communicate. What he proposes is a collection of empirical generalizations about language use and communication. If you want to teach someone how to use the word 'X', tell that person a goodly number of things from the following list of statements (about Xes). And so on for every word. The result is a large, unwieldy Berlitz-type manual. A traveler's or businessman's phrasebook differs from a standard textbook grammar in that its aim is to teach the kinds of responses needed to deal with typical situations. So the ordering and presentation of what is in the

phrasebook is a function not of the underlying, internal structure of the language but of situations that are likely to arise and the way they may unfold. The connection between syntax and semantics is broken. Sentences of widely disparate syntactical structure and complexity are juxtaposed. The structure is linear, following the direction of an imagined conversation. It does not equip the user with the ability to generate novel conversations. Thus a Berlitz-type manual, while useful in its own way, is not what Quine had in mind when he conceived of a translation manual as a theory of meaning. Putnam's theory of stereotypes is to the theory of meaning what knowledge of Berlitz French is to the Frenchman's knowledge of French. It has practical value, but lacks explanatory value.

Now to return to the matter of the meaning-belief distinction: Putnam's case against it was that holding to it deprives us of a theory of meaning that has explanatory power. Once we drop the distinction, such a theory can be had. But in the absence of such a theory, there is no reason to give up the meaning-belief distinction. So, given its prima facie evidence and appeal, the distinction should not be discarded. It is something that has to be dealt with directly in any theory of meaning. That is what Harman does in developing his conceptual role semantics, to which we now turn.

CHAPTER VI

HARMAN AND CONCEPTUAL ROLE

Harman holds that the central phenomenon of meaning is left unexplained if we rest satisfied with an account of how the meaning of expressions of natural language derives from speaker's meaning. This central phenomenon is the meaning of thoughts, the content of intentional states, the source from which expressions of natural language inherit their meaning. Consequently, Harman turns to the psychology of intentional states as a resource for semantics to exploit.

He looks to psychology also to provide a criterion of rationality different from Davidson's. With it, he sets out to tackle Quine's problem and isolate meaning without making Davidson's assumption of shared, true beliefs. The first move in Harman's approach is the claim that the psychology of Quine's native informants is not different from our own, that their minds work the way ours do. After all, we are members of the same species. The second move is the observation that humans normally say, not what is true, but what they believe (to be true). The third move is contained in the claim that the way the human mind works is, in principle, open to empirical investigation, and this over and above mere generalizations about sensory inputs and behavioral outputs. What makes other humans rational is

not, as Davidson would have it, that their beliefs are the same as ours; rather it is that their conceptual organization is the same as ours, that the beliefs that they have are formed in the same way that ours are formed, that their beliefs exhibit an internal coherence. The key to the new approach, which respects Quine's physicalism but not his behaviorism, is an empirically well-founded cognitive psychology of the states and processes that mediate between these inputs and outputs.

Given Harman's recourse to psychology, the structural similarity between his theory of meaning and Frege's may come as a surprise. However, it will prove useful, in order to get a picture of Harman's overall strategy, to explore its relationship to Frege's. This I shall do in the first section of the present chapter. In the second section, I shall draw from the relevant articles an account of Harman's theory. In the third section, I shall examine the yield of Harman's theory, comparing it in particular with that of Quine's. In the fourth section, I shall turn to criticisms of Harman's theory.

§1

It is not far from the mark to view Harman's semantics as Frege's theory of meaning turned on its head. Work in conceptual role semantics addresses an important pair of problems, how to provide an account of:

(1) the content of sentences and other linguistic expressions; and

(2) the content of thoughts, and in particular of beliefs.

Frege first recognized these problems as related, and attempted to tackle them in tandem. He begins by dividing the question of the content of linguistic expressions in two. Logicians from Aristotle to Gentzen have recognized that inferential relations hold among sentences of natural languages and have taken logic to be a theory of these relations. This view is implicit in Frege's characterization of logic as the effort to "discern the laws of truth."[1] Broadly speaking, these laws govern the inferential network into which sentences enter. This is not quite accurate on account of the ambiguity of natural languages. Taken in one sense, a sentence permits a certain range of inferences; taken in another sense, a fallacy may result. Making allowance for such ambiguity, Frege sees the goal of logic as the attempt to describe the inferential relations that hold among senses of sentences, or among sentences on a sense.

This role of sentences in inference is the first aspect of sentence content, and it is logic that furnishes the explanation of it. The sense of a sentence is its logical form, those properties in virtue of which the sentence's role in the inferential system is determined. As for other

---

[1]"The Thought: A Logical Inquiry," trans. A. M. & M. Quinton, Mind 65 (1956); reprinted in Essays on Frege, ed. E. D. Klemke (Urbana, Ill.: University of Illinois Press, 1968), 507.

linguistic expressions--predicates, proper names, connectives, etc.--their logical form or inferential content is a function of the logical form or inferential content of the sentences in which they occur.[2]

As well as having a role in inference, sentences are also bearers of truth and falsehood; they are about the world, or parts thereof. It is in fact from this property of sentences that semantics begins. Thus the primary semantic question is: how do sentences relate to the world? Frege's response again invokes sense. Sense, as well as incarnating logical form, determines reference. Thus both subquestions concerning the content of linguistic expressions--the nature of their inferential content and of their referential content--are answered by appeal to senses and the resources of logic.

Before turning to Frege's account of thought and belief, we need to recall his familiar distinction between thoughts (mental states, under which heading we may include propositional attitudes in general) and Thoughts (that is to say, Gedanken or propositions, which are none other than the senses of sentences). Frege's theory of propositional attitudes is literally that. Thoughts, beliefs and all propositional attitudes are attitudes toward, relations to, veritable propositions. The objects of thought and belief

---

[2]If the historical Frege is being stretched somewhat here, my purpose in this section of setting up a fairly standard intensionalist account of (1) and (2) for comparison with Harman's account is not thereby undermined.

are Thoughts. Thus in thinking we apprehend and consider objects that conform to a logico-deductive system. However, logical laws govern Thoughts, and apply only indirectly to (psychologically real) thoughts, functioning as an ideal, which actual human thought processes may follow, but fall short of:

> In one sense a law asserts what is; in the other it prescribes what ought to be. Only in the latter sense can the laws of logic be called "laws of thought": so far as they stipulate the way in which one ought to think.[3]

Finally, since objects of thought are senses of sentences, and sense is what determines reference, how thoughts are related to the world is a question that has already been answered and stands in no need of further explanation.

Frege's account of meaning accomplishes two fundamental tasks that are to be expected of a theory of meaning: for every meaningful expression of the language, there is a meaning assignable to it; and for any two expressions, there is a criterion by which they may be judged synonymous or not. On Frege's theory, to give the meaning of a linguistic expression will be to specify the properties of the expression that account for its role in inference and reference. Of course, for Frege these properties constitute the sense of the expression, since sense both incarnates logical form and determines reference. And two expressions

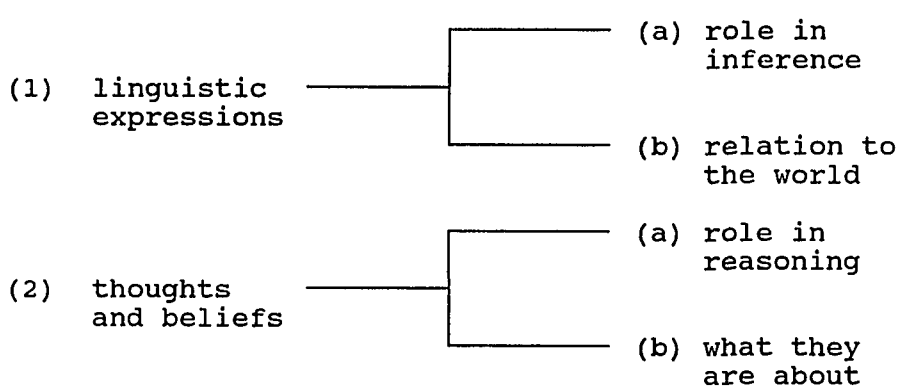
---

[3]The Basic Laws of Arithmetic: Exposition of the System, trans. and ed. Montgomery Furth, (Berkeley: University of California Press, 1964), 12.

will be synonymous just in case they have, not simply the same extension, but also the same sense. In addition, Frege's theory offers solutions to a whole series of problems involving reference, such as contingent identity sentences, failure of substitutivity in indirect discourse, denials of existence and non-referring singular terms.

§2

So much for the Fregean theory. My purpose in running briefly through it has been to see how, in recasting matters of thought content in terms of sentence content, Frege relies upon a parallel between the two. Just as sentences belong to an inferential network, so beliefs--and here I go beyond what is strictly Fregean--belong to a network that conveys evidential warrant, rather than truth, and also determines action. Moreover, thoughts and beliefs are, or purport to be, about the world, just as sentences are, or purport to be.



It is this framework, and not Frege's solution within it, that Harman retains. Like Quine, Harman rejects the

Fregean intensionalist approach because of qualms about the identity conditions of senses and about the account of the relation between sense and reference, which Frege takes to be straightforward while Harman finds it problematic and in need of further explanation. He also rejects Frege's treatment of human thought processes as either successful or failed attempts at deductive reasoning; according to Harman, reasoning "is rarely a conscious process"[4] while deductive logic serves as a tool that may be consciously employed in inference. Moreover, Frege's intensionalism presents a serious epistemological problem: How is it that humans come to know a language? How are we to conceive of the relation between human beings and senses? How do we grasp the objects of our propositional attitudes? Intensionalism stands in need of an epistemology.

Harman reverses Frege's strategy. Rather than subordinate thought content to sentence content, he takes the content of thoughts or mental representations as primary. Psychology supplants logic as the key to meaning.

But what kind of psychology is it that Harman invokes? It is a psychology that ascribes an ineliminable causal role in the explanation of human behavior to intentional states and mental processes. Intentional states are vehicles of informational content; mental processes are a matter of transformations upon this content. These transformations

---

[4]Thought (Princeton: Princeton University Press, 1973), 31.

require a medium, an internal system of mental representation, a "language of thought," whose formulae are the intentional mental states. Each mental representation is determined by the part that it plays in the total system; that part is its functional or conceptual role. In this psychology, to have a propositional attitude is to be in a particular computational state, and the computational states ascribed to a person are to be directly explicated as relations between the person and formulae in the person's internal representational system. We need to get a clear understanding of the three key elements of this account: intentional or representational mental states; their conceptual role in the internal system of mental representation; and the nature of the internal system itself.

A mental state is an internal state of a person, but not all internal states are representational. Being in pain, for instance, is not a representational state. Such internal states do not represent the external world as being one way rather than another; one's pain is neither true nor false; one is either in pain or one is not.[5] But other

---

[5]It may be objected that being in pain is also being in a representational state, that pain might represent physical trauma. On this view pain would be true if, say, there were a physical trauma and false if there were not:

> X is in pain iff there is a pain state p such that X bears relation R to p and p is true iff X is undergoing physical trauma T.

Suppose that every time X is undergoing physical trauma T, X is in pain. Pain state p drops out; it becomes otiose. Part of the very concept of a representation is the possibility of error, as McGinn has pointed out; that is to

internal states are representational; they have content. Believing that p, for example, is a representational or intentional state; my belief that p is either true or false; I may continue to have that belief regardless of its truth or falsity. While this psychology of intentional states breaks with Quine's behaviorism, which characterizes mental states as nothing more than dispositions to overt behavior, it does not violate his ontological strictures. Mental states may be token-identified with neural states, thus maintaining a physicalist ontology.

The content of each mental representation is determined by its conceptual or functional role, its place in a person's inner system of representation. Thus a conceptual role is a property of a mental representation, a particular kind of mental state, which in turn is a property of a person.

A person perceives goings-on in the world and engages in action. Between perception and action, mental processes occur. The best description of the processes linking perception to action, the best solution to the black-box

---

say, inaccurately representing its object does not preclude a representation's being a representation of that very object. (See the following chapter for more on this.) So we had better suppose that there are times when X is in pain but is not undergoing physical trauma T (assuming that there are no times when X is undergoing T but not in pain). In this case, there is no good reason to say that X is having a false pain; if we cleave to the physical basis of pain, then X's first-person pain-reporting loses its privilege and X is not in pain at all; or if we cleave to the experiential basis of pain, then X is in pain and the connection with trauma T is what is false.

problem this presents, attributes mental states to human beings, in particular, contentful mental states. Each mental state that a human is in at a given moment is related to perceptions past and present, to other non-perceptual mental states past and present and to actions present and future. These mental representations include not only beliefs, but also desires, hopes, doubts, etc. A new perception may engender a new belief. A new belief may lead to further new beliefs and desires; the elimination of old ones; and also actions. These connections to other states both actual and potential constitute the conceptual role of the belief in question. To characterize the conceptual role of a mental representation in isolation from its fellows is impossible. The enterprise is holistic. Once a description of the whole system has been given, each mental representation may be assigned its place in the scheme.

As for the nature of the internal medium of mental representations, Harman argues as early as his 1968 paper "Three Levels of Meaning"[6] and his 1970 paper "Language Learning"[7] that the language of thought is natural language, acquired and not innate. When we think in language, the language in which we think is the language that we speak. The inner system of representation of a native English speaker contains tokens of English expressions. This is not intended to deny infants the

---

[6]Journal of Philosophy 65 (1968): 590-602.

[7]Noûs 4 (1970): 33-43.

capacity to think. Rather, in learning a language one grafts onto a pre-linguistic system of representation a whole new system of thought. Natural language is a proper part of the system of representation. In addition to acquiring the ability to communicate, one also develops a new way of thinking, acquiring the potential to have thoughts one could never have had prior to learning the language. This view has important consequences.

First, learning a language is not a matter of learning a set of rules; rather it is coming to think in that language. Second, the whole Davidsonian problem of interpretation is a false one; communication is not a matter of decoding and encoding thoughts; to understand the speech of another is to recognize the sentence understood as having a particular deep structure. The meaning of sentences of the inner language is transparent; they do not require interpretation; and since the inner language is English there is no need either to worry about the interpretation of outer language. One simply recognizes a spoken sentence as a token of the same sentence type as an inner sentence token having a particular deep structure; the task is no more difficult than recognizing spoken and written tokens of the same sentence. This means that my mental representations are tokens of English sentences; they must therefore be analyzable as an ordered sequence of words having a deep structure; that they are neural states, and not acoustic waves or inscriptions, should not exclude them <u>a priori</u> from

being linguistic tokens. As Field suggests, they could be inscriptions on a tiny blackboard inside a person's head.[8]

Once the psychology is in place, how does it connect up with semantics? Harman takes the internal representational system to be the ultimate repository of linguistic meaning. The meaning of sentences used in communication is inherited from the meaning of sentences used in the language of thought, those internal formulae with which they are conventionally correlated.

> The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be used to express. . . . The contents of concepts and thoughts are determined by their functional role in a person's psychology. . . . 'Functional role' includes any special roles a concept may play in perception and in inference or reasoning, including practical reasoning that leads to action.[9]

Just as Quine holds that epistemology is "contained in natural science, as a chapter of psychology,"[10] so Harman has in mind a naturalized semantics, one continuous with the natural sciences, an empirical, not an a priori, enterprise. A functional account of an individual's psychology will describe an inner language game of belief-formation and decision-making, or, as Harman calls it, calculation, that involves the three elements of perception, inference and action.

---

[8]"Mental Representation," Erkenntnis 13 (1978): 27.

[9]"Conceptual Role Semantics," Notre Dame Journal of Formal Logic 23 (1982): 242.

[10]"Epistemology Naturalized," 83.

> Conceptual role semantics may be seen as a version of the theory that meaning is use, where the basic use of symbols is taken to be in calculation, not in communication, and where concepts are treated as symbols in a "language of thought."[11]

Thus the basic language game involves one person, a thinker, not a speaker/hearer. Speaker's meaning is derivative upon thinker's meaning. (This should not however be construed as leading to an approach along the lines of Grice or Searle; we shall see Harman's objections to those approaches shortly.) Processes of thought are the primary objects of semantic study. "The meaning of a linguistic expression is derived from its function in thought as determined by its place in [the thinker's] total conceptual scheme."[12] The theory of meaning is fundamentally a theory of the meaning or content of thought.

In sum, then, mental representations--beliefs, desires, hopes, doubts, etc.--enter into an inferential network governed by the laws of thought. These are the laws of psychologically real thought, not of Fregean Thought. It is the task, not of logic, but of cognitive psychology to describe the inferential relations that hold among them. Conceptual role is constituted by those properties in virtue of which the mental representation's function in the inferential system is determined.

The functional role of thoughts in calculation is one

---

[11]"Conceptual Role Semantics," 243.

[12]"Language, Thought and Communication," in Language, Mind and Knowledge, ed. Keith Gunderson, 283.

aspect of thought content. The second aspect of mental representations is their relation to the world. Just as for Frege sense determines reference, for Harman, conceptual role determines reference. For this to be achieved, conceptual role will have to be construed broadly, with functional relations extending beyond the domain of representation and into the outer world. In this way conceptual role will account for both aspects of meaning.

§3

What then is the yield of Harman's theory of meaning? In the first place, it responds directly to the kinds of considerations that led Frege to adopt his theory of senses. Where Frege invoked sense, Harman can offer solutions that turn upon conceptual role. The puzzle about contingent identity sentences is taken care of: 'Hesperus' and 'Phosphorus' have the same extension but different conceptual roles. The that-clause of a belief sentence determines the mental representation that is the immediate object of belief; substituting a coextensional expression within the that-clause will change the mental representation determined by the that-clause, unless it has the same conceptual role as the expression for which it is substituted; hence, it is plain why the truth value of the belief sentence may change. Denial of existence does not trap us into any untoward ontological commitment. The meaningfulness of 'Pegasus does not exist' commits us only to the existence of a mental representation, the one that is

typically correlated with the expression 'Pegasus'. Sentences containing non-referring singular terms are not thereby emptied of meaning, since they retain a conceptual role.

Of more importance to our investigation is how conceptual role semantics stacks up against Quine's theory. As for what is new, first, there will be a meaning assigned to each and every expression of the language: the conceptual role of the thought or concept with which it is typically correlated. Second, a pair of expressions will be synonymous when the conceptual roles of the mental representations with which they are conventionally associated are identical, or perhaps exhibit a sufficiently high degree of similarity. Third, mental representations, formulae in the language of thought, will supply objects of propositional attitudes. Fourth, the problem of incommensurability becomes tractable. Fifth, there will be a fact of the matter for translation to be right or wrong about, namely, conceptual role. As for what remains intact, there is no analytic-synthetic distinction; meaning remains a matter of evidence; no necessity, either de re or de dicto, is invoked; the holistic approach is mandatory; and reference remains inscrutable.

We need also to consider how Harman's approach fares against its competitors. It is Harman's contention that no other kind of theory, whether propositional, compositional, intention-based or speech-act theories, fully comes to grips

with the task of explaining how it is that noises and marks made by human beings come to be invested with meaning.

Simply to assign to sentences propositions as their meanings leaves wholly unexplained the nature of the relation between sentence tokens and their meanings. It should not remain a mystery. Semantics should satisfy certain epistemological demands, that how we know a language be understandable and plausible. The propositional account fails to fit in with causal theories of knowledge; so until another account is forthcoming, the propositional theories are lacking.

Compositional theories also fail to explain how noises and marks come to have meaning. The task that is accomplished by such theories is to show how the meanings of sentences are a function of the meanings of their parts. What it is for words to have meaning is left unaddressed. It is simply their lot in life to have the meanings they have which are graspable by humans. Just what it is that is grasped is left unaddressed.

Intention-based, Gricean theories take communication as the central use of language. What is communicated, the message, is a content. A theory of how a thought is expressed and understood leaves the content untreated; it sets out to explain how the content comes to be conveyed, how it is attached to those sounds. But anyone engaged in this project must recognize that linguistic meaning is inherited from thought content, which must be the more

basic.

A theory of the meaning of speech acts exhibits a double dependency. The performance of a speech act requires that the content of the speech act be communicated. Thus it presupposes a theory of meaning in communication, which in turn presupposes a theory of the meaning of thoughts.

Speech acts and messages inherit their meanings from the thoughts that are communicated in their performance. Compositional theories do not address the meaning of the primitives. Propositional theories fail to explain how noises and marks enter into their contingent relations with the propositions which are assigned to them as their meanings. Thus in all cases a theory of the meaning of thoughts is required. It is something in the production of those noises and marks that accounts for their having the meanings they do. A theory is needed to address this deficiency. It is such a theory that conceptual role semantics proposes.

§4

It is time now to see whether Harman's conceptual role semantics can make good on its claims. My criticisms fall into three areas: the relationship between semantics and psychology; the nature of the internal system of representation and the account of language learning that goes with it; and the account of reference.

The recourse to psychology naturally raises the charge of psychologism, that linguistics thereby loses its

autonomy. Harman has to make the case that what his conceptual role account provides falls within the domain of semantics. It is true that the functionalist psychology to which Harman makes appeal avoids reducing the semantic to the neurological. Semantic laws will thus remain independent of the laws governing the physical medium in which they are realized. However, why should psychological relations that hold among mental states count as semantic relations? How are we to distinguish between those relations that are properly semantic and those that are not?[13]

One of Quine's objections to Frege's intensionalist semantics was that, given the spuriousness of the analytic-synthetic distinction, there is no set of inferences, broader than the logical truths and narrower than the nomological, whose license is specifically semantic. Thus the range of inferences that a theory of meaning would have to account for collapses onto the logical truths. The risk for Harman is that the collapse would occur in the other direction, the semantical dissolving into the psychological. Harman may choose, however, to bite the bullet and accept that his theory does do away with the strictly semantical, but assert that it nevertheless accomplishes what we want of a theory of meaning.

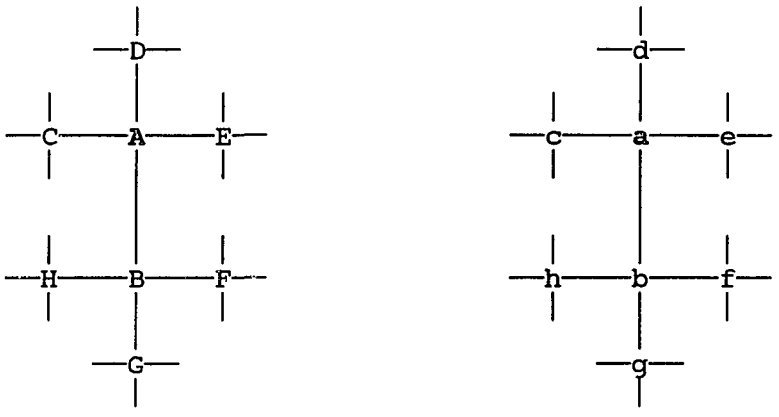
---

[13]Scott Soames has argued that making semantics a part of psychology means that facts that are intuitively non-linguistic (e.g., reaction times, error rates) cannot be excluded from semantics. See "Linguistics and Psychology," Linguistics and Philosophy 7 (1984): 155-180.

The radical holism that Harman adopts in defining conceptual role immediately gives rise to doubts about the theory's ability to meet those basic demands. The content of a mental representation is to be defined functionally in terms of its role in the total network of mental representation. To specify this content will require specifying the relations actual and potential to all other constituents of the network, an impossible task. And so, to give the meaning of any expression will be impossible. The problem is compounded by the fact that any change in the network entails a change in the conceptual role of every mental representation.

Synonymy suffers in similar fashion. For two mental representations to have the same content, the networks in which they occur must be identical, a demand that surely rules out not only inter-speaker synonymy but also intra-speaker synonymy over time. Lessening the demand to high degree of similarity, by, say, restricting the identity requirement to the proximate area of the network, is ad hoc and will not work in any case.

Let us assume that two networks are not identical but that we wish to specify synonymy conditions for mental representations A and a, which occur in sections of the two networks that look like this:



A has the same conceptual role (locally) as a iff:

AR_1B, AR_2C, AR_3D and AR_4E;
aR_1b, aR_2c, aR_3d and aR_4e;
B has the same conceptual role (locally) as b;
C has the same conceptual role (locally) as c;
D has the same conceptual role (locally) as d; and
E has the same conceptual role (locally) as e.

But B has the same conceptual role (locally) as b iff:

BR_5F, BR_6G, BR_7H and BR_8A;
bR_5f, bR_6g, bR_7h and bR_8a;
F has the same conceptual role (locally) as f;
G has the same conceptual role (locally) as g;
H has the same conceptual role (locally) as h; and
A has the same conceptual role (locally) as a.

But F has the same conceptual role (locally) as f . . .

Somewhere down the track, by hypothesis, we will find a difference between the two networks. No matter how distant that difference may be, its effect will still find its way back to the place we began. A and a will fail to have the same conceptual content. These two networks need not be in two different people. As soon as one of a person's beliefs changes or a new one is added, then the conceptual role of all the other elements of the network change, with the result that one new perception spells the end of intra-

speaker synonymy.

Thus two elementary demands upon a theory of meaning--that it give the meaning of all expressions and that it provide synonymy conditions--are not met. Harman may choose to give up synonymy, but in that case, the opacity of belief contexts looms as a problem for, since there are no synonyms, or even weaker substitutes, a criterion for substitution is lacking.

A second area of criticism focuses on Harman's interpretation of the research results in psychology. Jerry Fodor, for one, while concurring that "[t]he only psychological models of cognitive processes that seem even remotely plausible represent such processes as computational,"[14] draws a wholly different, non-empiricist lesson from cognitive psychology. Language learning, as well as concept learning, perception and choice, are, Fodor maintains, all matters of hypothesis formation and confirmation. The representational medium such hypotheses require bears many of the hallmarks of a language. Fodor's argument proceeds:

> Learning a language involves learning what the predicates of a language mean[, which] involves learning a determination of the extension of these predicates[, which] involves learning that they fall under certain rules. But one cannot learn that P falls under R unless one has a language [or language-like system] in which P and R can be represented. So one cannot learn a language unless one has a language [or language-like

---

[14]The Language of Thought (Cambridge, Mass.: Harvard University Press, 1975), 27.

system].[15]

Moreover, one cannot learn a language whose terms express semantic properties not expressed by the terms of some language, or language-like system, one is already able to use. The upshot of this argument is nativism: One cannot learn any predicate or concept whose properties one cannot already represent. All predicates and concepts must be either themselves innate or derivable from innate predicates and concepts. The inner representational system, "the language of thought," is not learnt.

If Harman is to maintain his empiricism, he has to offer an alternative account of language learning, one that will enable him to maintain the position that the "language of thought" is natural language, not an innate system of representation. This alternative is that "[l]earning an outer language involves the incorporation of that outer language into one's inner language."[16] But what incorporation involves Harman leaves largely unexplained. In his 1978 article "Is There Mental Representation?"[17] he does advance a number of considerations that are intended to lend support to the "incorporation view." When we learn a second language, there comes a time when we begin to "think

---

[15]The Language of Thought, 63-64.

[16]"Language Learning," 34.

[17]In Perception and Cognition: Issues in the Foundations of Psychology, vol. 9, Minnesota Studies in the Philosophy of Science, ed. C. Wade Savage (Minneapolis: University of Minnesota Press, 1978), 57-64.

in" that second language, rather than formulate our thoughts in our first language and then translate them. When we acquire new theoretical terminology, we develop new ways of thinking. Applying labels to elements of the environment enhances one's ability to perceive one's environment. We talk of "thinking out loud." But these common sense ideas may all be undercut.

When we "think in" French, it may be that we are encoding and decoding directly into and out of French, instead of engaging in a two-step process--from thoughts into English, and then from English into French, and the reverse. So it is by the relative ease of the straightforward one-step process that we are now struck. The new thoughts we come to have as a result of learning new terminology may be a simple result of having a whole set of new premises and axioms from which to draw new conclusions, not of having a restructured conceptual scheme. The fact that the employment of natural language labels increases the salience of environmental objects does not seem to cut one way or the other in the present dispute; it may simply enhance the accessibility of an internal concept. Stressing our ability to think out loud may simply underestimate our facility at encoding and decoding.

In the absence of any further indications beyond these heuristic observations on Harman's part about the mechanics of the incorporation view of language learning, it nevertheless seems clear that the incorporation view will

run into problems parallel to those of the British Empiricists. For rationalists, repeated exposure to, say, chairs, awakes in us a conscious awareness of the idea of a chair that was already innate in us. In this way, we come to recognize our sense impressions as tokens of a general idea.

For those empiricists who eschew innate ideas but allow general ideas, exposure to chairs produces simple ideas of chairs, upon which the mind performs an operation of abstraction, which yields the general idea of a chair. Harman cannot take this route because it means accepting that while our ideas are not innate, the relations between them are determined according to an innate scheme; and Harman has argued that the incorporation view involves new ways of thinking.

A different empiricist proposal comes from Hume, who eschews innate operations as well as innate ideas. Hume has the task of explaining how, without general or abstract ideas, we learn that chairs are chairs, as well as how ideas that are particular in their nature may yet be general in their representational function. General words furnish Hume with a solution. We hear the word 'chair' repeated in the presence of chairs, and thus all these particular ideas come to be associated with the word 'chair'. The word serves as the link between the particular ideas. Thus, general words replace general ideas. A particular idea of a chair can function generally when associated with the word 'chair'. A

problem arises, however, when we stop to ask how the individual occurrences of the word 'chair' come to be recognized as instances of the same word. In the case of chairs, there was recourse to the word 'chair', but with individual tokens of the word 'chair', no similar recourse is available; the word 'chair' is a general name for chairs, not for individual tokens of the word 'chair'. So a Humean approach runs into trouble, too.

At best, then, Harman has to count on future developments in psychology, and hope that they turn out to favor his incorporation view as opposed to the "code-breaking view" of his opponents.

I turn now to Harman's treatment of reference, the second aspect of thought content, which is to be generated from the first aspect, functional role within the internal representational system. His task appears difficult from the outset. The explanation of behavior at which a psychological theory aims appeals not to objects perceived or sought, but to the manner of representation of those objects. What makes behavior cohere is in the first place inside the head. The external world is at a remove from the domain of psychological explanation. The crux of the problem is that functional relations can be traced, so it seems, no further than the sensory periphery in one direction, and bodily movement in the other.

Harman rejects such solipsism. He maintains it is a mistake to assume that psychological theory is restricted to

what's in the head; conceptual role needs to be construed broadly, with functional relations extending beyond the domain of representation and into the outer world. In this fashion Harman seeks to account for "word"-world relations, and get around Putnam's Twin Earth examples.

There is good reason to think that he cannot. What he's after is a uniform functional account of relations inside and outside the head. Functional relations are independent of any particular realization; that is what they were designed to do. This is not the case for reference relations. Reference involves links to particular objects, not just to whatever object has a particular role. A functionalist account of conceptual organization provides something rather like a mental syntax. Syntax alone cannot generate semantic relations. So what Harman does is introduce semantics into the functionalist syntactic account under the guise of an extension of the range of psychology. Harman's position thereby loses much of its attractiveness, which lay in explaining content in terms of conceptual role without the built-in semantic element. Extending the notion of conceptual role to include semantic relations simply leaves reference unexplained. Moreover, the two aspects of content are conflated and obscured. In matters of reference you can't get something for nothing. Reference cannot be accounted for without invoking some relation to actual objects (this is hardly surprising); and that relation must be of a different order than relations that hold among

uninterpreted representations, whether of the same language or of a pair of languages.

Recognition of this independence of the two aspects of thought content has led a number of philosophers to develop dual-aspect versions of conceptual role semantics. It is their work that will be examined in the next chapter; only after that will a final assessment of conceptual role semantics be possible.

CHAPTER VII

FIELD AND DUAL-ASPECT SEMANTICS

We have noted how Frege divides sentence meaning into two components: role in inference and relation to the world. Having distinguished these two components, he proposes a unified solution to the problems of each. Sense is that which determines both logical form and reference. We have also seen how Harman follows Frege in recognizing two separate problems as well as in proposing a single solution, this time conceptual role, to them both.

Proponents of a dual-aspect[1] approach to conceptual role semantics recognize the Fregean distinction, and in fact cleave to it more faithfully than either Frege or Harman. The distinction, once made, is not to be discarded lightly. What dual-aspect theorists propose is two separate solutions to two separate problems. One solution they borrow from Harman and the other, to supplement it, they borrow from Davidson.

The eclecticism of dual-aspect theorists does not stop there. For a debt is owed also to Putnam, whose division of meaning into stereotype and extension presages the dual-

---

[1]Some philosophers employ the term 'two-factor theory' to refer to the same thing in order to distinguish it from theses in the philosophy of mind.

aspect theories we shall examine in this chapter. Stereotype is that part of meaning which is inside the head; and extension, that part of meaning which is outside the head. Putnam makes no effort to mold these separate elements into one whole (he is in fact more inclined to move in the opposite direction, breaking down stereotype even further). He recognizes that there are facts having to do with meaning that require two separate methods of explanation.

The ranks of the advocates of dual-aspect theories include (or have included) Hartry Field, Stephen Schiffer, Brian Loar, Colin McGinn and Ned Block.[2] These are the philosophers whose work I shall be considering in this chapter. I shall begin by laying out the most complete version of a dual-aspect theory--Field's--in order to see what it can do in contrast to Quine's and Harman's approaches. Then I shall examine in the subsequent sections the case for dual-aspect theories, both the negative case (consisting of charges against Harman's unified conceptual role semantics) and the positive case for supplementing

---

[2]Field, "Logic, Meaning, and Conceptual Role," Journal of Philosophy 74 (1977): 379-409; Schiffer, "Truth and the Theory of Content," in Meaning and Understanding, ed. H. Parret and J. Bouveresse (Berlin: Walter de Gruyter, 1981), 204-222; Loar, "Conceptual Role and Truth Conditions," Notre Dame Journal of Formal Logic 23 (1982): 272-283; McGinn, "The Structure of Content," in Thought and Object, ed. Andrew Woodfield (Oxford: Oxford University Press, 1982), 207-258; and Block, "Advertisements for a Semantics for Psychology," in Studies in the Philosophy of Mind, vol. 10, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein (Minneapolis: University of Minnesota Press, 1986), 615-678.

Harman's conceptual role semantics.

§1

Field's version of a dual-aspect theory consists of a conceptual role component and a truth-theoretic component. He follows Harman in adopting the hypothesis of an inner system of representation. However, Field casts the relations among the elements of the system and the conceptual roles of those elements in probabilistic terms, rather than in the causal terms that Harman employs and that other proponents of dual-aspect theories retain. Field shows how a subjective conditional probability function may be defined over the sentences of a person's inner representational system. This function is designed to capture the person's implicit assignment of probabilities to each of his or her internal sentences conditionally upon other sentences; the conceptual role of each internal sentence is determined by it. Given such a function, two sentences of a person's language of thought will have the same conceptual role if and only if they are equipollent with respect to that function, i.e., if and only if they have the same conditional probability with respect to the other sentences of the language. The immediate merit of Field's probabilistic version of conceptual role semantics is that it is the first precise account of what conceptual role is, precise enough in fact to yield the soundness and completeness of classical logic without appeal to truth.

Field, however, does not dismiss truth theory as Harman

does. While its place in Harman's theory is simply to serve as a disquotation device that supplies the conceptual roles of the truth-functional connectives, in Field's theory it assumes a place equal in importance to the account of conceptual role. The major break between Harman and Field occurs not in Field's reworking of the notion of conceptual role, which might be no more than an internecine dispute, but in his rejection of the claim that a theory of conceptual role is all there is to a theory of meaning.

Field's specific proposal for a second component is the modified version of a Tarski-style truth theory that he developed in his "Tarski's Theory of Truth."[3] In that place Field distinguishes three parts of Tarski's account of truth. The first part is the truth characterization, the recursive definition of truth in terms of a specified sequence of objects. It includes an inductive characterization of denotation for variables, constants and functions; satisfaction or truth for open sentences; and truth for closed sentences. The second part consists of set-theoretic definitions of denotation and satisfaction, which are required to supplement the truth definition so that there are no unexplained semantic terms. The third and most celebrated part is Convention T, which functions as the formal criterion of adequacy upon the truth definition.

This is the framework that Davidson adopted. Where Field departs from Davidson and Tarski is in his assessment

---

[3]Journal of Philosophy 69 (1972): 347-375.

of what the truth definition can do in the first part; and in the worth of the set-theoretic definitions in the second part. The first part defines truth relative to a sequence of objects and to a language, not truth plain and simple; and it assumes a notion of adequate translation. The truth definition, Field maintains, needs to be reworked and supplemented by a causal theory of reference, which has the real explanatory power that the empty, formal definitions of part two lack. So, although Field is taking over Davidson's idea, he once again diverges from his source. Field agrees with Harman that a truth theory that takes "'Beethoven wohnte in Deutschland' is true iff Beethoven lived in Germany" as its paradigm does no more than a manual that provides translations into a language previously understood. The paradigm he proposes in its place is:

> 'Beethoven wohnte in Deutschland' is true iff there are objects x and y and a relation R such that 'Beethoven' stands for x, 'Deutschland' stands for y, 'wohnte in' stands for R and x bears R to y.[4]

This paradigm requires that relations to extralinguistic entities, as well as the causal theory that explains them, be incorporated into one's semantics.

With the two components in place, we may now ask what this theory provides us. Having two components allows for two strategies in dealing with a problem. The first is to respond to a problem afflicting one component by invoking the other. For example, if the truth-theoretic component

---

[4]"Logic, Meaning, and Conceptual Role," 389.

cannot respond to the Frege problems, then the conceptual role component may be called upon. If the conceptual role component runs into trouble with the Twin Earth problems, the truth-theoretic component can deal with it. In this way, the dual-aspect theory inherits the yield of each of Harman's and Davidson's theories.

In particular, if Davidson's theory seems uncomfortably committed to accepting the T-sentence "'Grass is green' iff snow is white" as expressing the meaning of 'Grass is green', dual-aspect theories are not. It may appear that Field's move to probabilities does not avoid the problem Davidson encounters, since 'Snow is white' and 'Grass is green', for most speakers, will have the same subjective probability, namely, 1. However, it should be recalled that Field defines equipollence in terms of subjective conditional probability. So the subjective probability of the two sentences conditional upon, say, 'Below 0°C all objects develop a bluish tinge for which the human eye compensates' will differ; in the former instance, it will be less than 1, while in the latter it will remain at 1. Thus the two sentences are not equipollent, and so will, as we would want, differ in conceptual role, meaning, etc.

So, once it is recognized that truth conditions are not all there is to meaning and that conceptual role must also be taken into account, one is no longer tempted to substitute 'means that' for 'is true iff'. Conceptual role, McGinn suggests, is in fact what Davidson hinted at but

never developed when he saw the need for further constraints upon a Tarskian truth theory.[5] Conceptual role, moreover, responds to Davidson's call for a theory of meaning to be a theory of language understanding. The conceptual role factor explains how knowledge of meaning is integrated into behavior, something that a pure truth theory does not do.

The second strategy is to use the two components in tandem, to solve problems neither can deal with separately. In translating or interpreting, there are now two sets of constraints that serve to diminish indeterminacy: a sound translation will match both conceptual role and truth conditions. In dealing with propositional attitudes the two components can again work together. Field explicates the belief relation as a composite:

> X believes that p if and only if there is a sentence S such that X believes* S and S means that p,[6]

where S is a sentence of X's internal representational system, 'p' abbreviates a sentence of a natural language, and <u>believes*</u> is a dispositional relation (a matter of how the sentence is employed in reasoning, deliberating, etc.). The right-hand side of this equivalence treats belief as a relation to a mental representation that is characterized, in the first place, by its conceptual role, which

---

[5]"The Structure of Content," 240.

[6]"Mental Representation," 12. For reasons of exposition, Field introduces this claim in terms of sentences of outer language and possible worlds, but later discards them in favor of the theory he fully espouses, which is the one outlined here.

establishes its place in the overall system of mental representation. Furthermore, the second clause on the right-hand side is to be explicated by means of Field's referential version of a Tarski-style truth-theoretic semantics.

It may be objected against this explication of belief that there might be two sentences that express the same proposition, but that an individual will believe* one but not the other. This objection can receive either of two replies. First, Field can respond that the objection should be restated in the following form: two sentences may have the same referential truth conditions, but an individual will believe* one and not the other. This, of course, is not a problem; it is desirable that beliefs be individuated more finely than truth conditions alone permit.

The person raising the objection will respond that this is not the way the objection was meant to be taken. What was meant was indeed proposition, not truth conditions. To which a second reply may be made: What remains of the notion of proposition in Field's scheme is to be cast in terms of conceptual role. If two sentences have the same conceptual role, then when one is believed, so will the other. If not, then they do not have the same conceptual role.

As may be expected, Field's theory, drawing as it does upon the work of Davidson, Putnam and Harman, is further removed from Quine's base position. The above approach to

propositional attitudes is offered. While synonymy is not to be hoped for, its role in interpretation and translation is to be filled by sameness of conceptual role and of reference, which furnish constraints that limit the extent of indeterminacy. As I indicated in chapter V, the inclusion of the causal theory of reference entails the rejection of the inscrutability thesis and the acceptance of de re necessity. Although conceptual schemes may differ, the possibility of shared reference means that theory comparison is still possible. The only Quinean touchstones that remain intact are holism and the rejection of analyticity.

§2

It is time now to turn to the arguments for the theory and its various claims. The basic case for the inadequacy of Harman's conceptual role semantics turns upon the treatment of indexicals and Twin Earth examples. It is most clearly laid out by Block in "Advertisement for a Semantics for Psychology."[7] He asks us to consider two sentences:

(1) I am in danger of being run over.

(2) Ned Block is in danger of being run over.

If he were to utter (1) and (2), Block would, from one perspective, be saying the same thing, that is to say, expressing the same proposition, predicating the same property of the same object. His two utterances would have

---

[7]Pp. 618-627.

the same truth conditions.

However, consider the beliefs he would be expressing in uttering (1) and (2). The belief expressed by (1) would likely lead to evasive action of some sort. The belief expressed by (2) might not generate the same behavior, for Block might be under the mistaken impression that he is Napoleon. It would lead to evasive action only if supplemented by the belief Block would express by uttering:

(3) I am Ned Block.

Thus in uttering (1) and (2) Block would be expressing two beliefs that make different contributions to his behavior, two beliefs with different conceptual roles. From this perspective he would not be saying the same thing.

Even if Harman were to object to the view that utterances of (1) and (2) by Ned Block express the same proposition or have the same truth conditions, it nevertheless remains that conceptual role fails also to distinguish between Block's utterance of (1) and my utterance of the same sentence. The contribution of each expressed belief to the utterer's behavior would be the same; both of us would engage in evasive action. However, the truth conditions of the two utterances differ. And it is this difference that conceptual role is unable to capture.

The same point can be made by means of Putnam's Twin Earth examples. Putnam presupposes that my Twin Earth Doppelgänger and I are psychologically identical; so the

beliefs we express will have, by hypothesis, the same conceptual role. Yet they will have different truth conditions, since my beliefs are about H_2O and my Doppelgänger's are about XYZ. Again there are referential distinctions that conceptual role is unable to capture.

These examples reveal the two aspects of content. Block refers to them as narrow meaning--that aspect of content which is inside the head, contributes to the explanation of behavior and is to be explained in terms of conceptual role--and wide meaning--that aspect which is outside the head, has to do with reference and truth conditions and is to be addressed by a separate truth-theoretic component of a theory of meaning. Block warns us not to think, however, that narrow meaning is a part of wide meaning. So Loar's speaking of two dimensions is more felicitous: following Loar, the characterization of conceptual role occurs in a purely horizontal, "proof-theoretic" dimension, whereas truth conditions and references involve a vertical, "truth-theoretic" dimension.[8]

Harman's response has been to deny the distinction that Loar and Block seek to draw, and to maintain that conceptual role is not restricted to what is in the head. This in turn has led Loar and Block to develop a second line of criticism. They charge that the adjustments Harman makes to

---

[8]"Conceptual Role and Truth-Conditions," 274. See also his Mind and Meaning (Cambridge: Cambridge University Press, 1981), 57-58.

his theory in order to accommodate these kinds of objections are ad hoc.

Some of these adjustments are straight borrowings from Putnam. With the indexical and Twin Earth cases showing that in different contexts sameness of conceptual role does not preclude difference of reference, Harman borrows Putnam's solution, relativization to "normal context." Content will be conceptual role in a normal context. This maneuver may have its legitimate place in Putnam's theory, for his point is to reject psychological, "inside the head" theories of meaning and emphasize the causal-historical and social aspects of meaning. But Harman is trying to have it both ways. And the dual-aspect theorists do not so much reject this move as seek simply to be upfront about it. The reference relation is not solipsistically characterizable; it involves direct relations to entities outside the head. Harman tries to straddle the fence.

Their objection to Harman's borrowing of Putnam's notion of division of linguistic labor is on the same order. Given that I cannot tell a beech from an elm, how do I manage to be talking about elms rather than beeches when I use the word 'elm'? Putnam's solution, and Harman's, too, is to make appeal to the appropriate experts. The reference of my term is inherited from the reference of that term in the vocabulary of the tree experts in my linguistic community. However, as Loar points out, this leads to the problem how my term 'elm' and the experts' term 'elm' are

the same word, since they have, by hypothesis, different conceptual roles. It is perhaps overly stringent to demand sameness of conceptual role here. Perhaps instead a case should be made that a term in my inner language may borrow referential content from a term in someone else's when there is a high degree of similarity between the two terms' conceptual roles. But we have seen in the previous chapter how anything looser than sameness of conceptual role leads to a slippery slope. The only remaining recourse is to homophony, but that is, of course, a fallacy that begs the question whose answer we seek, for the terms of the problem for Harman are the same as those for the radical translator. A more general point should also be made. Conceptual role is again being stretched beyond its bounds to incorporate what properly belongs in the theory of reference. Better to distort neither theory and maintain the integrity of both by keeping them separate.

So Harman's extension of the purview of conceptual role is both ad hoc and obfuscatory. Moreover, by building in the thing represented, it runs afoul of the point McGinn makes about the necessary fallibility of representation.[9] McGinn, who subscribes to a version of dual-aspect semantics similar to Field's,[10] argues that the structures imposed

---

[9]"The Structure of Content," 210-211.

[10]McGinn proposes two subtheories. The first of these is a solipsistic theory of conceptual role cast in evidential terms and based, like Field's, on subjective probability. The second is a non-solipsistic Tarskian truth theory that may be supplemented by an account of the

upon our internal system of representation by conceptual role and truth-bearing role do not always coincide, that one role does not supervene on the other. Thus content is "inherently a hybrid of conceptually disparate elements."[11] Beyond the arguments based on consideration of the indexical and Twin Earth issues, he seeks to draw the same conclusion from a consideration of the nature, not of mental representation, but of representation in general.

Representations, he notes first, may succeed in their function even when they incompletely or improperly represent their object, for context may serve to establish the reference by making up for any lack of specificity on the part of the representation. Second, it is part of the very idea of a representation that it may represent more or less accurately, that it may even misrepresent; representation must be fallible. To account for these properties of representation, we need to call upon two distinct components of the content of a representation--what is represented and how it is represented. In the first case, any disparity between what is represented and how it is represented may be reconciled by context. In the second case, recognizing that

---

reference relation (as Field advocates) and is to be constrained by a condition of "extensional isomorphism," by which expression McGinn means that "the sentence mentioned and the sentence used in the theorems of the truth-theory should be composed of semantic primitives bearing the same extension in the same logical structure." ("The Structure of Content," 233.) His idea is to place a constraint on T-sentences that is stronger than material equivalence but weaker than synonymy.

[11]"The Structure of Content," 211.

what is represented is fixed independently of how it is represented permits sense to be made of the claim that a given representation is an inaccurate representation of its putative object rather than an accurate representation of some other object.

§3

All these considerations provide only a negative case for dual-aspect theories. They show simply that a unified conceptual role theory cannot do the job of accounting for truth conditions. As Loar says, "nothing in the description of at least some thoughts' or sentences' conceptual roles explains why we assign them references or truth-conditions."[12] Part of what one knows in mastering a language is truth conditions, and these involve environmental facts beyond the scope of conceptual role. But what is the positive case for the addition of a truth theory for one's inner system of representation? Given that conceptual role does not explain truth conditions and reference, what reason do we have to locate the truth theory at the level of the language of thought and not at the level of external language?

Field, Loar and Schiffer all give the same answer: reliability. In interpreting the speech of others, we seek not simply to gain knowledge about their state of mind, but also about the world. In so doing, we exploit the

---

[12]"Conceptual Role and Truth-Conditions," 274.

correlation between functional states and external states of affairs. This correlation is a contingent fact that does not follow from the nature of conceptual roles. And if our knowledge of the world were complete, we would not need to call upon it. But given the finitude of our knowledge, we need a reliability theory to apply to the beliefs of other people in forming our opinions about the world. Field puts his proposal this way:

> I think that the reason why we need to be able to apply semantic notions like truth and reference to the sentences that people believe and desire is that we hold the theory that people's beliefs are, in many circumstances, reliable indicators about the world; and the only way to state this theory is to use the notion of truth (and probably the notion of reference as well).[13]

Thus to obtain a theory of truth and reference for a person's inner system of representation, we construct a reliability theory, employing truth and reference as primitives, for that person. Then we use it to select the Tarski-style T-predicate that is the truth predicate for the person's inner system of representation.

In "Truth and the Theory of Content," Schiffer provides an even broader justification, arguing that the need for a Tarski-style T-predicate derives from its role in "the systematic exploitation of head-world reliability correlations"[14] not simply for knowledge about the world, but also to complete our psychology. Conceptual role, he

---

[13]"Mental Representation," 48.

[14]P. 217.

maintains, does not exhaust the domain of psychological explanation. He concedes to Field that if we had a conceptual role psychology plus easy access to all relevant environmental facts, we would have a complete explanation of behavior. However, without such access, we need to depend on the reliability of the beliefs of others to explain their behavior. That is to say, we need a reliability theory that incorporates a T-predicate. A T-predicate is of practical value. Moreover, there is no reason to expect there to be a unique T-predicate, i.e., one coinciding with the truth-predicate, because we are less than 100% reliable.

Thus it is from the area of communication that dual-aspect theorists draw their rationale for a truth theory for the inner system of representation. The first place to look for criticism of this rationale is Schiffer's recantation in Remnants of Meaning.[15] His counter-argument is that there is no good way of selecting from the range of available interpretations (T-predicates) of the language of thought the one that furnishes the truth conditions of the sentences of the language of thought. In the first place, people are unreliable, or at best reliable about some things and unreliable about others. So the formulation of a reliability theory requires an idealization, to the effect that, under optimal conditions, we human beings are reliable believers. The problem lies in the specification of these optimal conditions. We have no good idea what they might

---

[15](Cambridge, Mass.: MIT Press, 1987) ch. 4, 73-110.

be. Casting optimality as omniscience will not do. The human cognitive system is finite; and an idealization of it should not make it infinite. Schiffer goes on to assemble reasons why no reliability theory will be able to select out the truth-predicate for a person's language of thought. Underdetermination of theory suggests that there can be variance once we go beyond the observational. Both Tyler Burge (arthritis/shmarthritis)[16] and Saul Kripke (plus/quus)[17] have developed ways of generating counter-examples. Relativizing optimality to evolutionary survival needs does not do the trick either. Much of what is potentially knowable is of no evolutionary consequence. So again no sufficient constraint has been provided to assure that the truth-predicate will be selected.

McGinn also rejects reliability as a motivation for the second component. However, he retains the suggestion that communication is the source of a potential solution. He follows Field's lead on this, while all the same objecting to Field's argument that "(a) we want knowledge of the world (for obvious reasons), and (b) we take it that other people's utterances and beliefs are reliable ways of getting

---

[16]See "Individualism and the Mental," in Studies in Metaphysics, vol. 4, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein (Minneapolis: University of Minnesota Press, 1979), 73-121.

[17]See Wittgenstein on Rules and Private Language (Cambridge, Mass.: Harvard University Press, 1982), chs. 2 and 3.

such knowledge."[18] McGinn objects that these considerations are merely contingent. Field makes this a virtue, but McGinn points out that Field's position would crumble if either (a) or (b) were false.

Instead he chooses to focus attention not on the interpretation of speech, but on "the intentions with which assertions (and other kinds of speech act) are made."[19] Speakers seek not only to convey the solipsistic contents of their minds, but more importantly to convey information about the world. This reversal of point of view entails that reference is rooted in utterances and not in the language of thought.

McGinn notes that as a consequence of this position reference and truth belong properly to external language and not to the internal system of representation. "My amendment . . . does not seem applicable to beliefs at all, since belief-states are not actions performed with a point."[20] McGinn takes this in stride.

> I doubt that beliefs do have genuine referential properties in the manner of sentences--a belief's being <u>about</u> something is not the same as a sentence's <u>referring</u> to something.[21]

He situates the "aboutness" of belief somewhere between the "ofness" of perception and the reference of sentences in

---

[18]"The Structure of Content," 225.

[19]"The Structure of Content," 226.

[20]"The Structure of Content," 227.

[21]"The Structure of Content," 227.

terms of specificity and discrimination.

However, reference, we are asked to believe, is a product of an intention, which must be formulated in the inner system of representation. But if that system lacks the semantic specificity of external language, how is that referential intention to be formulated? We appear once again to be getting something for nothing. McGinn started off by saying that the referential content of external language is not specifiable in the language of thought. Yet it must be so specifiable if the language of thought is to be the source of it. So McGinn's amendment to Field has not worked.

Moreover, McGinn's maneuver may be viewed as a concession by someone committed to the single-aspect approach, who can shift reference outside the theory of meaning proper and into the theory of communication, while maintaining that the contribution of meaning to truth conditions is simply the determination of a function from context to reference. Such a function does not require the establishment of relations extending outside the head.

In conclusion I return to Field's probabilistic proposal for the conceptual role component. Field states that

> conceptual role semantics, though it deals nicely with questions of intraspeaker synonymy, cannot properly answer questions about interspeaker synonymy or about relations between language and the world.[22]

---

[22]"Logic, Meaning, and Conceptual Role," 380.

The problem with this statement is that even the modest claim regarding intra-speaker synonymy is not borne out. We saw with Harman's theory that not only does inter-speaker synonymy go by the board but so does intra-speaker synonymy over time. On Field's version, too, the conceptual role component fails in that conceptual role is defined relative to a subjective probability function for an individual at a time. For different individuals, or for the same individual at different times, the probability function will differ, the psychological properties, states and relations it is intended to capture will have evolved. Recall that two sentences of a person's language of thought will have the same conceptual role if and only if their conditional probability with respect to <u>each and every</u> sentence is identical. So no two individuals can have the same thought unless they happen to be described by identical probability functions, and no individual speaker can have the same thought over time. It seems clear that any approach built on the idea of evidential warrant will engender the same result. As for Field's claim that sameness of reference can fill in for synonymy, the failure of the second component puts an end to that option.

So rather than resolve the problems inherent in the conceptual role and truth-theoretic approaches by combining the two, the dual-aspect approach simply compounds them.

CHAPTER VIII

FINAL THOUGHTS

Quine's work in semantics has sought to prove a pair of results:

(A) Intensionalism is untenable and, as a result, an extensionalist base for the construction of a semantic theory is the only one available; and

(B) There is a (low) upper limit on any theory constructed on that base.

Each of the theories that have been examined in the preceding five chapters starts from the position that while (A) is true, (B) is false. I have focused exclusively on the second issue, seeking to prove that (B) is in fact true, and hence that Word and Object places an upper limit upon any extensionalist theory of meaning. In the first section of this chapter, I wish to reinforce that result. Then, in the second section, I shall consider the bearing of this result on (A). If Quine is right about (A), then Word and Object sets an upper limit, not just on extensionalist semantic theories, but also on semantic theory of any kind.

§1

Each of the strategies we have considered seeks to lift the ceiling that Quine has imposed upon the theory of

meaning by expanding the extensionalist base in one direction or another, but not by building upon an intensionalist base. The failure of these individual strategies does not automatically rule out the truth of their position. It might still be possible to maintain the truth of (A) and the falsity of (B) by claiming that:

(i) further tinkering with one or another of these strategies may yet work; or

(ii) a new strategy that is nevertheless extensionalist may be devised.

However, the particular flaws in these strategies are so fundamental that (i) is ruled out, and a diagnosis of their general faults rules out (ii).

Conceptual role strategies come in two forms, Harman's single-aspect version and the dual-aspect version espoused at various times by a number of different philosophers. Harman's unsatisfactory explanation of the relation of mental representations to their external reference (involving such issues as indexicals and Twin Earth cases) led us to seek a solution among the dual-aspect approaches. But they were unable to justify the attribution of reference to mental representations rather than to expressions of external, natural languages. More fundamental than the problem provided by the referential component is the problem with the notion of conceptual role. If meaning is a matter of conceptual role, then there is no accounting for inter-speaker synonymy or even intra-speaker synonymy over time,

for any trivial psychological difference or change results in a difference or change of meaning, and not just for a particular expression but, on account of the holistic nature of the theory, for every expression. The problem remains the same even if we switch to Field's probabilistic account of conceptual role. And it also remains if we weaken the demand for synonymy to high degree of similarity in meaning. Basically both the single- and dual-aspect strategies are too tied to individual psychology. They assign too fine-grained a structure to semantic content.

Putnam's strategy also involves two parts: a stereotype component and a reference component. The problem with his causal-social account of the latter is that it seems to serve better as a theory of those factors that determine change in reference over time; it does not provide an explanation of the reference relation itself but requires a theory of speaker's reference, which is not within its own purview. As for the stereotype component, psychological evidence runs against the thesis that what the speaker knows is encompassed in a stereotype. Moreover, the problem of determining what belongs to the stereotype and what does not reproduces the problem of determining which truths are analytic and which are not; and there is no guarantee that those properties belonging to a stereotype will not fail to be properties that, on an intuitive basis, a translation ought to preserve. In addition, there is the problem of establishing a general account of the similar-to-the-

stereotype relation. In any case, Putnam's strategy, even if it were successful, does not represent a real departure from the Quinean position that semantic and empirical information cannot in general be separated.

Lewis employs three key notions in his modal approach: possible world, actuality and counterpart. We have seen that his theory runs into trouble on all three counts. First of all, unlike Kripke, who simply stipulates worlds (no contradiction in the stipulations means the world is a possible one), Lewis specifies worlds qualitatively and must then determine whether they are possible; but he lacks a general procedure for determining whether a world is possible. Second, to treat the word 'actual' as an indexical referring to this world, Lewis needs to say what kind of thing this world is; but this task is left undone; there is no scientific theory of what this world is (let alone what a possible world is)[1] and common sense, even assuming that it does not reject Lewis's modal realism out of hand, is too imprecise for his needs. And third, as for counterpart theory, adopted by Lewis to avoid the problem of trans-world identity, what it lacks is a general scale to measure degrees of comparative similarity; even admitting that the problem of measuring comparative similarity is no worse across worlds than within a world, it nevertheless remains just as troublesome.

---

[1]A physical theory of all the things in the world need not contain the predicate 'is a world'.

The initial attraction of Davidson's program lay in the simplicity of the proposal to employ Tarski's apparatus in semantics. A claim that meaning is simply a matter of truth conditions has a certain ring to it: Get rid of meanings, just deal with truth conditions; and that we know how to do because of Tarski's work. But the subsequent elaboration of Davidson's strategy has been almost a series of retreats, in which the initial insight has given way to the recognition of the increasing importance of the empirical constraints that need to be added to the formal constraints imposed by the adoption or the truth-theoretic approach. Davidson has gone so far as to adopt Lewis's way of casting the theory in terms of two components, one formal and the other empirical: on the one hand, a grammar, a formal system empty of empirical content that creates a pattern or structure; on the other, a theory of interpretation that gives it empirical content.[2] As the theory of interpretation has come to shoulder more and more of the burden of the theory, the truth-theoretic component has receded into the background. Davidson claims that the latter is still the most vital part of the theory. But that is a vague claim. As truth theory diminishes in importance, Davidson runs the risk that those critics, such as Harman, who have claimed that truth theory does nothing more than fix the meanings of the logical constants may be right. If so, then the truth-theoretic component does not ascribe enough semantic

---

[2]See "The Structure and Content of Truth," 279-328.

structure to the language. Further formal constraints are needed, not just the empirical constraints of the theory of interpretation. But the theory of interpretation, which is required to determine the set of T-sentences for a language, also founders because it turns upon the assumption that the interpreter must attribute true beliefs, namely, our beliefs, to the person whose speech is being interpreted. Davidson's methodological argument for this assumption is unsound.

The question that remains to be addressed in this section is why the failure of these strategies should lead us to believe that no new extensionalist strategy might work. One path is to argue that together they cover a broad range of strategies. But I think that would be a mistake. In truth, these theories draw on one or the other of a pair of resources, the only kinds of resources available: empirical/material or formal/grammatical. To the first kind of theory belong Putnam's realist approach as well as the psychological approach of the conceptual role theorists, both single- and dual-aspect. To the second kind of theory belong the truth-theoretic and possible world strategies. These are the two ways of introducing more structure into the fabric, mesh or network of language, to employ Quine's metaphors.

The problem with the recourse to empirical resources is that in the one case it introduces too little structure and in the other too much. What structure Putnam's account

assigns to a language is ad hoc, lacking in systematicity and no more than what Quine allows; in the last analysis, Putnam's stereotype theory may be a prescription for the kind of empirical lexicography that Quine dismisses in "Two Dogmas" as mere "reports upon usage."[3] As for the psychologically-based theories, whether they ascribe a causal or a probabilistic structure to the inner system of representation, they necessarily attribute too much structure to the language; they make distinctions that are too fine-grained, which results in the synonymy problems. Thus these empirical approaches either overshoot the mark or undershoot it.

The problem with the recourse to formal resources--grammars based on truth theory or possible-world theory--is that if the only evidence for the theory is T-sentences, in the former instance, or what we may call T-in-a-possible-world-sentences, in the latter, then they will necessarily ascribe too little structure, or too coarse a grain, and so fail to determine properly semantic relations. A theory that does not have resources at least as powerful as intensionalism cannot succeed in creating semantic relations out of weaker ones.

§2

Now, if Quine is right about the truth of (B), is he also right about (A)? There are three basic positions that

---

[3]P. 25.

may be taken against intensions:

(1) that they do not exist;

(2) that theories employing them are vacuous and fail to be explanatory; and

(3) that there is nothing for them to explain.

The truth of (1) cannot be decided a priori. By Quine's own standard of ontological commitment, if there is a sound theory that is the best theory of the facts, then, like it or not, we are committed to the existence of those entities whose existence is postulated by that theory. It does no good for the extensionalist to protest that the fewer the abstract objects we accept, the better off we are. Some well-behaved abstract objects are ontologically acceptable because a sound, well-formed theory, e.g., set theory, requires them. So the truth of (1) depends upon that of (2) and (3).

The intensionalist can successfully respond to (2). An intensionalist theory can meet the demand for formal explanatory adequacy. For example, such a theory could be constructed after the fashion of Davidson's or Lewis's theory. Both Davidson and Lewis divide a theory of meaning into two components, a grammar and theory of interpretation. It should be noted that this division is not the same as the one propounded by dual-aspect theorists or by Putnam, who divide up purely empirical content into internal and external before formulating empirical theories of each kind. In contrast, Davidson and Lewis divide up the theory of

meaning into: a formal component empty of empirical content, a grammar of a possible language; and an empirical/material component that interprets that grammar by identifying it as the grammar of a particular natural language. On this approach, there is no need to place any empirical constraints on the first component, whose purpose is to specify the structure of a (possible) language; formal adequacy is all that is required of it. What this strategy does is loosen the methodological constraints upon the theory-formulation component by exporting them to the theory-confirmation component.

So there is nothing to stop a linguist from constructing an abstract, formal system that contains all the usual intensionalist structure--analyticity, ambiguity, redundancy, synonymy, etc.--along the lines of Davidson's truth theory. With this difference, however, that it will provide recursive definitions of not just a truth-predicate, but also an analyticity-predicate, an ambiguity-predicate, a redundancy-predicate, a synonymy-relation, etc. To be more precise, since these are as-yet uninterpreted predicates: not just a T-predicate, but also what we may refer to as an An-predicate, an Am-predicate, an R-predicate, an S-relation, etc. It will then be the job of the theory of interpretation to identify them with the truth predicate, analyticity predicate, etc. for some actual language. To do this the theory of interpretation will have to establish the T-sentences, An-sentences, Am-sentences, R-sentences,

S-sentences, etc. of that language. For it is those sentences that will constitute the evidence for the theory.

Here the discussion arrives at position (3). At this point, we may have an intensionalist grammar that has the form of an explanatory theory. The only hitch, Davidson, Lewis and Quine may point out, is that there is no confirming evidence. The intensionalist requires evidence in the form of semantic intuitions of analyticity, ambiguity, etc.--not just attitudes of holding true or assenting and dissenting--and there are just no such properties. Judgments of that sort are bogus and unacceptable as confirming evidence for any theory or grammar; semantic theory is useless because there is nothing to explain, no explananda, no confirming evidence.

Now, Quine allows judgments of assent and dissent and judgments about T-sentences as well, although not intuitions about possible worlds, judgments about T-in-a-possible-world-sentences. Why not admit more evidence? The behaviorism of Word and Object cannot be assumed. So there must be some other basis for Quine's position. The objection is that the kind of evidence needed by the intensionalist is judgments of a second-order, about sentences. And sentential attitudes, apart from the privileged ones of assent, dissent and holding true, which Tarski has shown to be admissible, suffer from opacity. Getting rid of this opacity requires propositions, synonymy, analyticity and that whole circle of notions that was

repudiated in "Two Dogmas."

But if there seems to be a circularity here, it is not at all clear which side is guilty of it. On the one hand, it may be claimed that: semantic theory is useless because the judgments that serve as the confirming evidence or the explananda of the theory are bogus, and the evidence or explananda are bogus because no sense can be made of them by any non-vacuous theory. But on the other hand, the counter-claim can be made that: the theory is not useless because there is evidence to confirm it and explananda it serves to explain, and the evidence and explananda are not bogus since a working formal theory is available. Evidence and theory go hand in hand. One camp rejects both; the other accepts both. The intensionalist and the extensionalist are at a stand-off over (3).

However this stand-off is resolved, what is now clear is that the intensionalist's most formidable opposition remains Quine, for the attempt to build an extensionalist theory of meaning of greater scope than Quine's is bound to fail.

BIBLIOGRAPHY

Armstrong, Sharon Lee, Lila R. Gleitman and Henry Gleitman. "What Some Concepts Might Not Be." Cognition 13 (1983): 263-308.

Block, Ned. "Advertisements for a Semantics for Psychology." In Studies in the Philosophy of Mind. Vol. 10, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein, 615-678. Minneapolis: University of Minnesota Press, 1986.

_____, ed. Readings in Philosophy of Psychology. 2 vols. Cambridge, Mass.: Harvard University Press, 1980.

Burge, Tyler. "Individualism and the Mental." In Studies in Metaphysics. Vol. 4, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein, 73-121. Minneapolis: University of Minnesota Press, 1979.

Carnap, Rudolf. Meaning and Necessity. 2nd ed. Chicago: University of Chicago Press, 1956.

Cartwright, Richard. Philosophical Essays. Cambridge, Mass.: MIT Press, 1987.

Davidson, Donald. "Theories of Meaning and Learnable Languages." In Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science, ed. Yehoshua Bar-Hillel. Amsterdam: North-Holland, 1965. Reprinted in Inquiries into Truth and Interpretation, 3-15.

_____. "Truth and Meaning." Synthèse 17 (1967): 304-323. Reprinted in Inquiries into Truth and Interpretation, 17-36.

_____. "On Saying That." Synthèse 19 (1968-69): 130-146. Reprinted in Inquiries into Truth and Interpretation, 93-108.

_____. "Semantics for Natural Language." In Linguaggi nella Società e nella Tecnica. Milan: Edizioni di

Comunità, 1970. Reprinted in Inquiries into Truth and Interpretation, 55-64.

_____. "Radical Interpretation." In Dialectica 27 (1973): 313-328. Reprinted in Inquiries into Truth and Interpretation, 125-139.

_____. "Belief and the Basis of Meaning." Synthèse 27 (1974): 309-323. Reprinted in Inquiries into Truth and Interpretation, 141-154.

_____. "Thought and Talk." In Mind and Language: Wolfson College Lectures 1974, ed. Samuel Guttenplan, 7-23. Oxford: Oxford University Press, 1975. Reprinted in Inquiries into Truth and Interpretation, 155-170.

_____. "The Method of Truth in Metaphysics." In Studies in the Philosophy of Language. Vol. 2, Midwest Studies in Philosophy, ed. Peter French, Theodore Uehling, Jr., and Howard Wettstein. Minneapolis: University of Minnesota Press, 1977. Reprinted in Inquiries into Truth and Interpretation, 199-214.

_____. "Reality without Reference." In Dialectica 31 (1977): 247-253. Reprinted in Inquiries into Truth and Interpretation, 215-225.

_____. Inquiries into Truth and Interpretation. Oxford: Clarendon Press, 1984.

_____. "The Structure and Content of Truth." Journal of Philosophy 87 (1990): 279-328.

Davidson, Donald, and Jaakko Hintikka, eds. Words and Objections: Essays on the Work of W. V. Quine. Dordrecht, Holland: Reidel, 1975.

Davis, Steven. "Translational Indeterminacy and Private Worlds." Philosophical Studies 18 (1967): 38-45.

Descartes, René. Meditations on First Philosophy. In Philosophical Writings, trans. and ed. Elizabeth Anscombe and Peter Geach. London: Thomas Nelson, 1954.

Field, Hartry. "Tarski's Theory of Truth." Journal of Philosophy 69 (1972): 347-375.

_____. "Theory Change and the Indeterminacy of Reference." Journal of Philosophy 70 (1973): 462-481.

_____. "Quine and the Correspondence Theory." Philosophical Review 83 (1974): 200-228.

_____. "Conventionalism and Instrumentalism in Semantics." Noûs 9 (1975): 375-406.

_____. "Logic, Meaning, and Conceptual Role." Journal of Philosophy 74 (1977): 379-409.

_____. "Mental Representation." Erkenntnis 13 (1978): 9-61.

Fodor, Jerry. The Language of Thought. Cambridge, Mass.: Harvard University Press, 1975.

_____. Representations: Philosophical Essays on the Foundations of Cognitive Science. Cambridge, Mass.: MIT Press, 1981.

Frege, Gottlob. "On Sense and Meaning." Translated by Max Black. In Translations from the Philosophical Writings of Gottlob Frege, ed. Peter Geach and Max Black, 56-78. 3rd ed. Totowa, N.J.: Rowman & Littlefield, 1980.

_____. The Basic Laws of Arithmetic: Exposition of the System, trans. and ed. Montgomery Furth. Berkeley: University of California Press, 1964.

_____. "The Thought: A Logical Inquiry." Translated by A. M. & M. Quinton. In Essays on Frege, ed. E. D. Klemke, 507-535. Urbana: University of Illinois Press, 1968.

Gunderson, Keith, ed. Language, Mind and Knowledge. Vol. 7, Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press, 1975.

Harman, Gilbert. "Quine on Meaning and Existence, I," Review of Metaphysics 21 (1967): 124-151.

_____. "Three Levels of Meaning." Journal of Philosophy 65 (1968): 590-602.

_____. "Language Learning." Noûs 4 (1970): 33-43.

_____. Thought. Princeton, N.J.: Princeton University Press, 1973.

_____. "Meaning and Semantics." In Semantics and Philosophy, ed. Munitz and Unger, 1-16.

_____. Rev. of Meaning, by Stephen Schiffer. In Journal of Philosophy 71 (1974): 224-229.

_____. "Language, Thought and Communication." In Language, Mind and Knowledge, ed. Keith Gunderson, 270-298.

_____. "An Introduction to 'Translation and Meaning.'" In Words and Objections, ed. Davidson and Hintikka, 14-26.

_____. "Is There Mental Representation?" In Perception and Cognition: Issues in the Foundations of Psychology. Vol. 9, Minnesota Studies in the Philosophy of Science, ed. C. Wade Savage, 57-64. Minneapolis: University of Minnesota Press, 1978.

_____. "Conceptual Role Semantics." Notre Dame Journal of Formal Logic 23 (1982): 242-256.

_____. Change in View. Cambridge, Mass.: MIT Press, 1986.

Haugeland, John, ed. Mind Design: Philosophy, Psychology, Artificial Intelligence. Cambridge, Mass.: MIT Press, 1981.

Hume, David. A Treatise of Human Nature. 2nd ed. Oxford: Clarendon Press, 1978.

Kant, Immanuel. Critique of Pure Reason. Translated by Norman Kemp Smith. New York: St. Martin's Press, 1965.

Katz, Jerrold J. Philosophy of Language. New York: Harper & Row, 1966.

_____. "Logic and Language: An Examination of Recent Criticisms of Intensionalism." In Language, Mind and Knowledge, ed. Keith Gunderson, 36-130.

_____. Language and Other Abstract Objects. Totowa, N.J.: Rowman and Littlefield, 1981.

_____. "The Naturalistic Fallacy in Semantics." In Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson, ed. Ernest LePore, 59-91.

_____. "The Refutation of Indeterminacy." Journal of Philosophy 85 (1988): 227-252.

Kripke, Saul. Naming and Necessity. Cambridge, Mass.: Harvard University Press, 1980.

_____. Wittgenstein on Rules and Private Language. Cambridge, Mass.: Harvard University Press, 1982.

LePore, Ernest, ed. Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson. Oxford: Basil Blackwell, 1986.

LePore, Ernest, and Barry Loewer. "Dual Aspect Semantics." In New Directions in Semantics, ed. Ernest LePore, 83-112. London: Academic Press, 1987.

Lewis, David K. "Counterpart Theory and Quantified Modal Logic," Journal of Philosophy 65 (1968): 113-126. Reprinted in Philosophical Papers, vol. 1, 26-46.

_____. Convention. Cambridge, Mass.: Harvard University Press, 1969.

_____. "General Semantics." In Semantics of Natural Language, ed. Gilbert Harman and Donald Davidson, 169-218. Dordrecht, Holland: Reidel, 1972. Reprinted in Philosophical Papers, vol. 1, 189-232.

_____. Counterfactuals. Oxford: Basil Blackwell, 1973.

_____. "'Tensions." In Semantics and Philosophy, ed. Munitz and Unger. Reprinted in Philosophical Papers, vol. 1, 250-260.

_____. "Radical Interpretation." Synthèse 27 (1974): 331-344. Reprinted in Philosophical Papers, vol. 1, 108-121.

_____. Philosophical Papers. Vol. 1. Oxford: Oxford University Press, 1983.

_____. On the Plurality of Worlds. Oxford: Basil Blackwell, 1986.

Loar, Brian. Mind and Meaning. Cambridge: Cambridge University Press, 1981.

_____. "Conceptual Role and Truth Conditions." Notre Dame Journal of Formal Logic 23 (1982): 272-283.

Loewer, Barry. "The Role of Conceptual Role Semantics." Notre Dame Journal of Philosophy 23 (1982): 305-315.

Lycan, William. "Semantics and Methodological Solipsism." In Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson, ed. Ernest LePore, 245-261.

Mates, Benson. "Synonymity." In University of California Publications in Philosophy 25 (1950): 201-226. Reprinted in Semantics and the Philosophy of Language, ed. Leonard Linsky, 111-136. Urbana: University of Illinois Press, 1952.

McGinn, Colin. "The Structure of Content." In Thought and Object, ed. Andrew Woodfield, 207-258. Oxford: Oxford University Press, 1982.

Munitz, Milton K., ed. Identity and Individuation. New York: New York University Press, 1971.

Munitz, Milton K., and Peter Unger, eds. Semantics and Philosophy. New York: New York University Press, 1974.

Perry, John. "The Problem of the Essential Indexical." Noûs 13 (1979): 3-21.

Putnam, Hilary. "The Analytic and the Synthetic." In Scientific Explanation, Space, and Time. Vol. 3, Minnesota Studies in the Philosophy of Science, ed. Herbert Feigl and Grover Maxwell. Minneapolis: University of Minnesota Press, 1962. Reprinted in Mind, Language and Reality, 33-69.

_____. "Is Semantics Possible?" In Languages, Belief and Metaphysics. Vol. 1, Contemporary Philosophic Thought, ed. H. Kiefer and M. Munitz. New York: SUNY Press, 1970. Reprinted in Mind, Language and Reality, 139-152.

_____. "The Meaning of 'Meaning'." In Language, Mind and Knowledge, ed. Keith Gunderson, 131-193. Reprinted in Mind, Language and Reality, 215-271.

_____. Mathematics, Matter and Method. Vol. 1, Philosophical Papers. Cambridge: Cambridge University Press, 1975.

_____. Mind, Language and Reality. Vol. 2, Philosophical Papers. Cambridge: Cambridge University Press, 1975.

_____. Meaning and the Moral Sciences. London: Routledge & Kegan Paul, 1978.

_____. Reason, Truth and History. Cambridge: Cambridge University Press, 1981.

_____. Realism and Reason. Vol. 3, Philosophical Papers. Cambridge: Cambridge University Press, 1983.

_____. "Possibility and Necessity." In Realism and Reason, 46-68.

_____. "Reference and Truth." In Realism and Reason, 69-86.

Quine, Willard Van Orman. "Two Dogmas of Empiricism." Philosophical Review 60 (1951): 20-43. Reprinted in From a Logical Point of View, 20-46.

_____. "Three Grades of Modal Involvement." In Proceedings of the XIth International Congress of Philosophy. Amsterdam: North-Holland, 1953. Reprinted in The Ways of Paradox, 158-176.

_____. From a Logical Point of View. 2nd ed. Cambridge, Mass.: Harvard University Press, 1961.

_____. "Quantifiers and Propositional Attitudes." Journal of Philosophy 53 (1956): 177-187. Reprinted in The Ways of Paradox and Other Essays, 185-196.

_____. Word and Object. Cambridge, Mass.: MIT Press, 1960.

_____. The Ways of Paradox and Other Essays. New York: 1966. Enlarged ed. Cambridge, Mass.: Harvard University Press, 1976.

_____. "Ontological Relativity." Journal of Philosophy 65 (1968): 185-212. Reprinted in Ontological Relativity and Other Essays, 26-68.

_____. Ontological Relativity and Other Essays. New York: Columbia University Press, 1969.

_____. "Epistemology Naturalized." In Ontological Relativity and Other Essays, 69-90.

_____. "Reply to Chomsky." In Words and Objections, ed. Davidson and Hintikka, 302-303.

_____. "On the Reasons for the Indeterminacy of Translation." Journal of Philosophy 67 (1970): 178-183.

_____. Review of Identity and Individuation, ed. Milton K. Munitz. In Journal of Philosophy 69 (1972): 488-497.

_____. "On Empirically Equivalent Systems of the World." Erkenntnis 9 (1975): 313-328.

_____. "Facts of the Matter." In American Philosophy from Edwards to Quine, ed. Robert W. Shahan and Kenneth R. Merrill, 176-183. Norman, Okla.: University of Oklahoma Press, 1977.

_____. Theories and Things. Cambridge, Mass.: Harvard University Press, 1981.

_____. "Indeterminacy of Translation Again." Journal of Philosophy 84 (1987): 5-10.

Rorty, Richard. "Indeterminacy of Translation and of Truth." Synthèse 27 (1972): 443-462.

Rosch, Eleanor. "On the Internal Structure of Perceptual and Semantic Categories." In Cognitive Development and the Acquisition of Language, ed. Timothy E. Moore, 111-144. New York: Academic Press, 1973.

Rosenberg, Jay. "Synonymy and the Epistemology of Linguistics." Inquiry 10 (1967): 405-420.

Schick, Karl. "Indeterminacy of Translation." Journal of Philosophy 69 (1972): 818-832.

Schiffer, Stephen. Meaning. New York: Oxford University Press, 1972.

_____. "Truth and the Theory of Content." In Meaning and Understanding, ed. H. Parret and J. Bouveresse, 204-222. Berlin: Walter de Gruyter, 1981.

_____. "Compositional Semantics and Language Understanding." In Philosophical Grounds of Rationality: Intentions, Grounds, Ends, ed. Richard Grandy and Richard Warner, 175-205. Oxford: Clarendon Press, 1986.

_____. Remnants of Meaning. Cambridge, Mass.: MIT Press, 1987.

Searle, John. "Minds, Brains, and Programs." In Mind Design: Philosophy, Psychology, Artificial Intelligence, ed. John Haugeland, 282-306. Montgomery, Vermont: Bradford Books, 1981.

Sellars, Wilfrid. "Some Reflections on Language Games." In Science, Perception, and Reality. London: Routledge & Kegan Paul, 1963.

Soames, Scott. "Linguistics and Psychology." Linguistics and Philosophy 7 (1984): 155-180.

Stich, Stephen. "Davidson's Semantic Program." Canadian Journal of Philosophy 6 (1976): 201-227.

Tarski, Alfred. "The Concept of Truth in Formalized Languages." In Logic, Semantics, Metamathematics, trans. J. H. Woodger, 152-278. Oxford: Clarendon Press, 1956.