"Donald Davidson," in *Key Thinkers in the Philosophy of Language*, ed. Barry Lee, Continuum Press, 2011.

PREPRINT

Donald Davidson

Kirk Ludwig

1. Introduction

Donald Davidson was one of the most influential philosophers working in the theory of meaning in the latter half of the twentieth century. He can be credited with one of the few genuinely novel approaches to the theory of meaning that emerged in that period, namely, the program of truth-theoretic semantics and its integration into the theory of radical interpretation.

The first of Davidson's two sub-projects is his famous though controversial proposal, advanced originally in "Truth and Meaning" (1967), to use a Tarski-style truth theory (a theory that determines truth conditions for all of the sentences in a language on the basis of axioms for each semantically primitive expression in it) to do the work of a compositional meaning theory (one that gives the meaning of each of a language's sentences on the basis of the meanings of its parts and their mode of combination). This builds on Tarski's pioneering work on how to define a truth predicate for a formal language[1] (Tarski, 1944, 1983). Central to this was Tarski's Convention T, which requires that an adequate truth theory have as theorems all sentences of the form (T)

(T) *s* is T iff *p*

in which (i) 'is T' is a truth predicate in the metalanguage (the language in which the theory is expressed) for the object language (the language the theory is about), (ii) '*s*' is replaced by a description of an object language sentence as composed out of its significant parts (a structural description, for short), and (iii) '*p*' is

[1] A formal language has a precisely defined set of well-formed formulas generated recursively with formation rules operating over a set of primitive symbols, as for the languages of symbolic logic. It may be partially or fully interpreted.

replaced by a metalanguage sentence that translates it. In contrast to Tarski, Davidson was not interested in formal but in natural languages, and he was not interested in defining a truth predicate but in adapting the kind of truth theory Tarski showed how to construct for natural languages, and in using a predicate known antecedently to express the concept of truth, in order to further the goals of a compositional meaning theory. Davidson's second sub-project is his proposal, inspired by Quine's project of radical translation (1960), to investigate the concepts of the theory of meaning by reflection on how a radical interpreter could confirm a truth theory for a speaker's language ("Radical Interpretation" (1973)). The sub-projects represent two parts of a single enterprise: that of explaining what it is for words to mean what they do (2001b, p. xiii). The first aims to help explain how we understand complex expressions on the basis of their significant parts. The second aims to shed light on meaning more generally by showing how one can understand another speaker with whom one does not (initially) share a language, on the basis of his behavior in relation to others and his environment. This relates facts about meaning to the context of communication in which they get their purpose and to the evidence on the basis of which we must perforce interpret others, and so relates meaning to the more fundamental facts upon which it supervenes.

The suggestion that a truth theory can be used as a meaning theory has sometimes been met with incredulity. Ostensibly, a truth theory states only conditions under which a sentence is true, and though truth is determined by meaning and how the world is, specifying conditions under which a sentence is true seems to guarantee no insight into its meaning. For example, while (S)

(S) 'Snow is white' is true iff grass is green

specifies a condition under which 'Snow is white' is true, this yields no insight into the meaning of 'Snow is white'. It has therefore been suggested that Davidson must really have intended either the reduction of meaning to some (perhaps special, "strong") notion of truth conditions (Burge, 1992, pp. 20-21; Horwich, 2005, p. 4 & ch. 8) or the replacement of the traditional pursuit of a meaning theory with a successor project on the grounds that the notion of meaning is too confused for systematic investigation (Chihara, 1975;

Cummins, 2002; Glock, 2003, p. 142 ff.; Katz, 1982, pp. 183-185; Soames, 1992, 2008; Stich, 1976). We'll see that each of these suggestions rests on a misunderstanding of how Davidson intends a truth theory to aid pursuit of a meaning theory.

In the following, we approach the twin themes of Davidson's work in the theory of meaning by looking at the motivations for the introduction of a truth theory as a vehicle for a meaning theory, and the influences on Davidson's choice of the position of the radical interpreter as the most fundamental from which to investigate the concepts of the theory of meaning.

2. Compositionality

Davidson's work in the theory of meaning starts with the observation that natural languages are compositional, in the sense that they admit of a division into semantically primitive expressions and semantically complex expressions that are understood on the basis their primitive constituents and mode of combination. For example, the two sentences 'John loves Mary' and 'Mary loves John' are different in meaning but are understood on the basis of the same primitive vocabulary items and rules for their combination.

Davidson argued that natural languages are compositional because they have an infinity of nonsynonymous sentences but are mastered by finite beings. As Davidson puts it (1965, p. 9— all citations to page numbers are to reprints of articles in Davidson's collected papers, as indicated in the references), "we do not at some point suddenly acquire an ability to intuit the meanings of sentences on no rule at all" and "each new item of vocabulary, or new grammatical rule, takes some [minimum] finite time to be learned." On this basis, Davidson introduces a requirement on an acceptable meaning theory for a natural language:

> I propose what seems to me clearly to be a necessary feature of a learnable language: it must be possible to give a constructive account of the meaning of the sentences in the language. Such an account I call a theory of meaning for the language, and I suggest that a theory of meaning that conflicts with this condition … cannot be a theory of a natural language; and if it ignores this condition, it fails to deal with something central to the concept of a language. (1965, p. 3)

The observation that natural languages are compositional seems straightforward, but places important constraints on meaning theories for natural languages. An adequate theory must exhibit expressions as falling into two classes and show how expressions falling into the category of semantically complex expressions are understood on the basis of understanding primitives and rules for their combination. A piecemeal approach, that does not take into account the full range of uses of expressions in sentences, will not meet the adequacy condition. And an account of any range of discourse that requires an infinite number of semantical primitives cannot be a correct account of a natural language.[2]

3. Criticism of the appeal to meanings as entities

A venerable tradition in accounting for the compositionality of natural languages, one that stretches back to Frege (see chapter 1), involves assigning to every expression in the language a meaning that is grasped by anyone who understands it, and which determines the expression's extensional properties (relative to the way the world is), that is, its referent, its extension, or its truth value (as it is a name, predicate or sentence). The compositionality of natural languages then is expressed as the thesis that the meanings of complex expressions are functions of the meanings of their components.

Davidson took a dim view of this, and understanding why is crucial to understanding how he thought of the role of a truth theory in a meaning theory. To see the difficulty, suppose we associate a meaning with 'Theatetus', say, Theatetus, and with the predicate 'flies', say, the property of flying. The trouble is that by itself this gives us no insight into the meaning of 'Theatetus flies'. So far as anything we have said goes, it might as well be a list: Theatetus, the property of flying. It is natural to say that concatenation is itself semantically significant, and that it means *instantiates*. Consistency requires we assign concatenation a meaning, the relation of instantiation. Now we simply have a longer list: Theatetus, the relation of

[2] Davidson thought, for example, that both Frege's and Carnap's treatment of belief sentences violated this constraint. Davidson's interest in the problems of compositionality was sparked initially by reflection on Carnap's treatment of belief sentences (Davidson, 1963).

instantiation, the property of flying. We want to say that 'Theatetus' and 'flies' do not both function as proper names. But this can't be determined by what object each term is associated with. We need a rule that shows how to interpret their combination, and which exhibits their different contributions to the formation of an expression of a different semantic type than either of them, which is evaluable as true or false.

Davidson illustrates the point with a simple reference theory.[3] Take a language fragment L whose primitive vocabulary consists of the names 'Marie' and 'Jean', which refer to Mary and to John, and the functor 'La mère de'. A singular term in the language consists of a name or the concatenation of 'La mère de' with any singular term. This generates an infinite set of referring terms. We want a theory that gives the reference of any expression in this infinite language fragment. Suppose we assign a function *f* to 'La mère de' and say that for any singular term α, the referent of 'La mère de' $\frown\alpha$ is the value of *f* given the referent of α as argument (where '$\frown$' is the concatenation symbol). This yields *no insight* into the referent of any complex expression. To make the function scrutable, we must add that the value of *f* for any *x* is the mother of *x*. But then it is clear that the assignment of the function is not what is doing the work, but the use of 'the mother of' in saying what the combination of 'La mère de' with a singular term refers to. We might as well eliminate the middleman: for any singular term α, the referent of 'La mère de' $\frown\alpha$ is the mother of the referent of α.

There are two points to take away from this. The first is that in explaining the function of a complex expression on the basis of its parts one must give a rule. The second is that it is neither sufficient nor necessary for this to assign an entity to every expression. For in our simple theory, we found that assigning an entity to 'La mère de' did not help with the work of the theory and that the work of the theory could be done without assigning any entity to it. So even in the context just of a reference theory, it is clear that it is misguided to think that the work of the theory is either advanced by, or requires, assigning an entity to

[3] I change the example to a fragment of French to set the stage for generalizing and extending the point.

every component expression of a complex term. This point, as we will see, extends to developing a meaning theory for a full language.

A simple extension turns our reference theory into a meaning theory. Although Davidson did not take this step with the example, it parallels the step he takes later with the truth theory. Seeing it in this simplified context will help appreciate how it works for the truth theory. We start by stating a criterion of adequacy on a reference theory: it should entail, for each object language singular term, a theorem of the form '*s* refers to *t*' where '*s*' is a structural description of an object language singular term and '*t*' is replaced by a metalanguage term that translates it. Call this Convention R. For our sample theory,

1. 'Marie' refers in L to Mary
2. 'Jean' refers in L to John
3. For any singular term α in L, 'La mère de' $^\frown\alpha$ refers in L to the mother of what α refers to in L.

Convention R is satisfied because in the first two axioms the referents of the object language names are given using metalanguage names that translate them, and in the third axiom a metalanguage functor the same in meaning as the object language functor is used to give the rule. Any theory that satisfies this kind of constraint on its *axioms* satisfies (we will say) Convention A. The point of introducing Convention A is that it applies at the level of axioms, unlike R, which applies at the level of theorems, and this is important to a point we will make in a moment. Satisfying Convention A (relative to an appropriate class of proofs) suffices for the theory to satisfy Convention R, though not in general vice versa. If the theory satisfies Convention R, then—and this is the payoff—from a theorem of the form [r] we can infer a corresponding instance of [m],

[r] *s* refers in L to *t*.
[m] *s* means in L *t*.

because [m] is true if '*t*' translate *s*. Thus, we specify the meaning of each expression in our infinite language fragment on the basis of a finite number of rules attaching to its primitive expressions. Moreover, a step-by-step proof of a reference theorem shows how the meaning of each expression contributes to

fixing the referent of the complex expression, for given that the theory meets Convention A (this is why Convention A turns out to be important), we reflect in the axioms the meanings of the primitive expressions and in the proofs their contributions in virtue of meaning to fixing the referents of complex referring terms. And as noted in the previous paragraph, this is accomplished without the need to assign an entity to every expression, and, specifically, our complex term forming device, 'La mère de'.

We identify a theory of meaning for a language with *the body of knowledge that puts us in a position to understand each expression in it*. The meaning theory then is not identical with the reference theory. For we have to know more than what the reference theory states to interpret each object language expression. We have to know (i) what the axioms are; (ii) what they state so-specified; (iii) that the theory satisfies Convention A, and so Convention R; (iv) a canonical proof procedure from axioms to the canonical theorems in virtue of which the theory satisfies Convention R; and (v) the inference rule that takes us from canonical theorems of the form [r] to [m].

Before extending these ideas to a whole language, we should review a final proposal Davidson considers for a compositional meaning theory that does not assign meanings to every expression, but does assign meanings to sentences. The rejection of this proposal is important for understanding what he does next. The proposal is to extend the idea in the reference theory sketched to a whole language by treating sentences as referring to their meanings and predicate expressions like '*x* is red' as functioning like 'the mother of x'. For example, we might give the following axiom for 'est rouge'.

For any singular term α, $\alpha \frown$'est rouge' refers to <u>the referent of α is red</u>.

Instantiated to 'Marie', we can infer

'Marie est rouge' refers to <u>Marie is red</u>.

We may replace 'refers to' if we like with 'means'. Davidson seeks to scotch this proposal with a famous argument dubbed the Slingshot by Barwise and Perry (1981), which he attributes to Frege. It is too involved

to go into here. Its conclusion is that if sentences are treated as singular terms referring to their meanings, on plausible assumptions, it follows that all sentences alike in truth value refer to the same thing, and hence have the same meaning—an intolerable result. The argument, though, is unsuccessful, for it either equivocates or begs the question (Lepore & Ludwig, 2005, pp. 49-55). Despite this, there is little to recommend the view that sentences refer to anything, let alone what they mean, and the obstacles in the way to a workable theory along the lines sketched are non-trivial. If there is any other way to get the desired result, it would be preferable.

4. The proposal to use a truth theory in pursuit of a meaning theory

Davidson concludes:

[a] What analogy [with the reference theory] demands is a theory that has as consequences all sentences of the form '*s* means *m*' where '*s*' is replaced by a structural description of a sentence and '*m*' is replaced by a singular term that refers to the meaning of that sentence; a theory, moreover, that provides an effective method for arriving at the meaning of an arbitrary sentence structurally described. … Paradoxically, the one thing that meanings do not seem to do is oil the wheels of a theory of meaning—at least as long as we require of such a theory that it non-trivially give the meaning of every sentence in the language. (1967, pp. 20-21)

However, this leaves us with the problem proving theorems of the form '*s* means that *p*' from axioms attaching to primitives where we treat neither '*p*' nor 'that *p*' as a referring term. Davidson suggests that, "in wrestling with the logic of the apparently non-extensional 'means that' we will encounter problems as hard as, or perhaps identical with, the problems our theory is out to solve" (p. 22). For we can substitute after 'means that' only on the basis of synonymy, which looks to require us to already have a meaning theory (for the metalanguage) of the sort we want to develop for the object language, which would force a regress. It is in the face of this that Davidson makes the proposal to use a truth theory (1967, pp. 22-23).

[b] The only way I know to deal with this difficulty is simple, and radical. … The theory will have done its work if it provides, for every sentence *s* of the language under study, a matching sentence (to replace '*p*') that, in some way yet to be made clear, 'gives the meaning' of *s*. One obvious candidate for matching sentence is just *s* itself, if the object language is contained in the metalanguage; otherwise a translation of *s* in the metalanguage. As a final bold step, let us try treating the position occupied by '*p*' extensionally: to implement this, sweep away the obscure

'means that', provide the sentence that replaces '*p*' with a proper sentential connective, and supply the description that replaces '*s*' with its own predicate. The plausible result is

(*T*) *s* is *T* if and only if *p*.

What we require of a theory of meaning for a language *L* is that without appeal to any (further) semantical notions it place enough restrictions on the predicate 'is *T*' to entail all sentences got from schema *T* when '*s*' is replaced by a structural description of a sentence of *L* and '*p*' by that sentence.

[c] … it is clear that the sentences to which the predicate 'is *T*' applies will be just the true sentences of *L*, for the condition we have placed on satisfactory theories of meaning is in essence Tarski's Convention *T* that tests the adequacy of a formal semantical definition of truth.

To see how this is a clever pursuit of the initial project by indirect means, we can rephrase the design problem for an adequate meaning theory described in [a] in a more general way:

formulate a theory that has as consequences all sentences of the form '*s* … *p*', where '*s*' is replaced by a structural description of sentence and '*p*' by a metalanguage sentence that *gives the meaning* of that sentence; a theory, moreover, that provides an effective method for arriving at the meaning of an arbitrary sentence structurally described.

If a theory issued in true theorems of the form (M)

(M) *s* means that *p*

on the basis of axioms for primitive expressions in the language, then it would satisfy this criterion. This essentially comes to matching an object language sentence *s* with a metalanguage sentence '*p*' in use that translates *s*. The difficulties in formulating such a theory appear formidable. Davidson's insight was to see that a truth theory that meets Convention T would in effect satisfy the criterion because it requires the truth theory to entail every instance of the (T)-schema in which '*s*' is replaced by a structural description of an object language sentence and '*p*' by a metalanguage sentence that translates it. This is the same relation that must hold between *s* and '*p*' for true instances of (M). We could, then, restate Convention T as the requirement that the truth theory entail all instances of (T) for which the corresponding instances of (M) are true. Thus, if an instance of (T) is one of the theorems of a truth theory in virtue of which it meets

Convention T, then the corresponding instance of (M) yields an explicit statement of the meaning of the object language sentence.

To see this in more detail, consider a simple truth theory [T] for a fragment of a language L without quantifiers or context sensitivity. (A language is context sensitive if it contains elements whose contribution to truth conditions is determined relative to a context of utterance, as in the case of tense, whose contribution is relative to the time of utterance, and pronouns like 'I', which refers to the person using it, and 'that', which refers to what the user demonstrates in using it. 'I like that', for example, expresses the speaker's partiality at the time of utterance toward what he demonstrates in using 'that'.)

1. 'Marie' refers in L to Mary
2. 'Jean' refers in L to John
3. For any name α, $\alpha \frown$'dort' is true in L iff what α refers to is sleeping.
4. For any names α, β, $\alpha \frown$'aime' $\frown \beta$ is true in L iff what α refers to loves what β refers to.
5. For any sentence φ, 'Ce n'est pas le cas que' $\frown \varphi$ is true in L iff it is not the case that φ is true in L.
6. For any sentences φ, ψ, $\varphi \frown$'et' $\frown \psi$ is true in L iff φ is true in L and ψ is true in L.

Suppose we know, as for the reference theory, that each axiom uses a metalanguage term that translates the object language expression. So 'Mary' translates 'Marie', 'John' translates 'Jean', 'is sleeping' translates 'dort', etc. If a truth theory meets this condition, then (parallel to the reference theory) it meets Convention A[4] and is *interpretive*. If it meets Convention A, it meets Convention T. For example, a theorem of this theory is (T*), from which we can infer (M*).

(T*) 'Ce n'est pas le cas que Jean aime Marie' is true in L iff it is not the case that John loves Mary.

(M*) 'Ce n'est pas le cas que Jean aime Marie' means in L that it is not the case that John loves Mary.

Call a T-sentence like (T*), where the metalanguage sentence on the right translates the sentence for which it gives truth conditions, *interpretive*. Importantly, the theory does more than issue in interpretive theorems. Because it satisfies Convention A, the proofs of the relevant theorems reveal at each stage the

[4] As in the case of the reference theory, placing this requirement on the axioms goes beyond anything Davidson said, but it is implicit in the goal of providing a compositional meaning theory by way of a truth theory. See (Lepore & Ludwig, 2005, pp. 71-74, 2007, pp. 34-39).

contribution of each primitive object language term, by way of the contribution that invoking the axiom for it makes, to fixing the interpretive truth conditions for the object language sentence, in virtue of the meaning of the object language term. That is, the theorem's proof reveals the compositional structure of the object language sentence for which interpretive truth conditions are given. We have, thus, "a constructive account of the meaning of the sentences in the language," which provides, in Dummett's apt phrase, "a theoretical representation of a practical ability" (1993, p. 36).[5]

As for our sample reference theory, what enables us to interpret object language sentences includes more than the truth theory itself. To infer (M)-sentences from (T)-sentences we have to know that the theory satisfies Convention T; to use the proofs to reveal compositional structure we have to know that the axioms satisfy Convention A, and know a canonical proof procedure that will terminate in the appropriate canonical theorems. Thus, the meaning theory is not the truth theory per se but an appropriate body of knowledge about it.

We have illustrated the idea with respect to a small fragment of a language without quantifiers or context sensitivity. Convention A and Convention T can be modified to suit a context sensitive language with quantifiers, and the conceptual points carry over straightforwardly.[6]

5. The extended project and the reduction and replacement interpretations

The project as presented is an enlightened pursuit of a compositional meaning theory for natural language without the expedient of assigning meanings as entities to every expression. Why has Davidson sometimes been taken, then, to have been aiming to reduce meaning to truth conditions or to reject giving a meaning

[5] This is not to claim that competence is realized by propositional knowledge of the theory. In "A Nice Derangement of Epitaphs" (1986, p. 438), e.g., Davidson says, "To say that an explicit theory for interpreting a speaker is a model of the interpreter's linguistic competence is not to suggest that the interpreter knows any such theory…They are rather claims about what must be said to give a satisfactory description of the competence of the interpreter."

[6] See chapter 5 of (Lepore & Ludwig, 2005) for a basic discussion of the modifications needed to accommodate context sensitivity and see (Lepore & Ludwig, 2007) chapters 2-3 for a discussion of quantifiers, and chapters 4-11 for a discussion of context sensitive referring terms and tense.

theory, on the grounds that it is irremediably confused, in favor of a best successor project? The answer lies in a proposal that he makes after proposing that a truth theory can be used as a vehicle for a compositional meaning theory, which represents his pursuit of what we can call the extended project, in contrast with the initial project of providing a compositional meaning theory. The idea of the extended project is to place substantive constraints on a truth theory that ensure that it satisfies Convention T. The point of this extended project is to reveal connections between meaning facts and facts that are not about meaning as such. If we could place non-semantic constraints on a truth theory that sufficed for it to satisfy Convention T, we could claim to have shown something important about what grounds facts about meaning. Given the goal, it is clear that this is not a rejection of giving a meaning theory, but an attempt to give one in an illuminating way.

Why should this give rise to a misinterpretation of Davidson? The two primary reasons are, first, that in his earliest presentation of the idea, Davidson had not cleanly drawn the distinction between the truth theory and the meaning theory, and, second, that at precisely the point at which Davidson suggests a truth theory can be used to get around the problems facing the direct approach, he likewise shifts, without explicitly indicating that he is doing so, his attention from the initial to the extended project. For he thought that he saw an opportunity, once we transition to a truth theory for a language with context sensitive expressions, and especially demonstratives, of placing a simple substantive constraint on a truth theory that would suffice for it to satisfy Convention T, namely, that it be extensionally adequate—that is, that it simply be a true theory. Thus, it can seem that Davidson was suggesting that a correct truth theory is on its own a theory of meaning. But since a truth theory does not say anything about meaning, and merely matching a mentioned object language sentence with a used metalanguage sentence correlated in truth value does not give the meaning of the object language sentence, it can seem that what he says cannot be taken at face value. Surely he must be suggesting that meanings can be reduced to truth conditions, or that talk of meaning is so obscure that it must be replaced in serious discussions by a theory that deals with

more tractable concepts and offers the prospect of systematic investigations, much as explanations of illness in terms of evil spirits have given way to the germ theory of disease. By now it is clear that both of these interpretations involve serious mistakes about how the truth theory is supposed to play its role. First, the meaning theory is not the truth theory per se, but a certain body of knowledge about the truth theory. Second, the suggestion that an extensionally adequate truth theory would serve as a meaning theory, Davidson's first suggestion, was conditional on the supposition that extensional adequacy would suffice for the theory to meet Convention T.

Why did Davidson think extensional adequacy was enough? For example, as we noted earlier, although (S) is true, it hardly helps us understand 'Snow is white'.

(S) 'Snow is white' is true iff grass is green

There were two connected ideas. The first was that once we had moved to a context sensitive language, we would have a very fined-grained test for the adequacy of a theory in the need to accommodate the truth conditions of sentences containing demonstratives and other context sensitive expressions. The second was that we would treat the theory as an empirical theory, which would then be responsible for all actual and potential utterances of speakers of the language, and so would be responsible for getting right anything anyone might say about any object. Thus, it would have to issue in the right truth conditions for sentences such as 'That is grass', 'That is snow', 'That is white', and 'That is green' in application to any object. (S) would not survive this test, for the theory that generated it would also predict for example that an utterance of 'that is white' would be true of something iff 'that is green' was.

However, this initial case for taking extensional adequacy to be enough was found to be inadequate (Foster, 1976; Loar, 1976). Accommodating demonstratives would rule out such T-sentences as (S), but would not rule out *all* non-interpretive T-sentences. We could modify any predicate axiom of a truth theory by adding to the truth conditions for it an eternally true sentence, such as 'the earth moves' or '2+2=4'. For example,

> For any speaker *s*, time *t*, for any name α, α ⌢'dort' is true as used by *s* at *t* in L iff the referent of α as used by *s* at *t* is sleeping at *t* and the earth moves.

This would generate a non-interpretive canonical theorem when instantiated to, e.g., 'Marie' because the corresponding M-sentence is false. When this objection was raised Davidson returned to the problem of specifying a substantive constraint on a truth theory that would enable it to be used to interpret object language sentences in the context of the project of radical interpretation, which is the subject of the next section. That Davidson tried to improve on his initial suggestion shows decisively that the reduction and replacement interpretations are incorrect, for on either account he would have had no reason to change is his view about the adequacy of his proposal.

6. What is the project of radical interpretation?

Radical interpretation is a successor to Quine's project of radical translation (see chapter 5). The radical translator approaches the task of understanding another speaker without any prior knowledge of the speaker's meanings or attitudes. He restricts himself to the speaker's dispositions to verbal behavior in response to stimulus in constructing a translation manual for the speaker's language, and thus isolates the empirical content of a theory of translation. Translation manuals alike in empirical content were judged to capture all the meaning facts which there were. In "Epistemology Naturalized" (1969), Quine explains the ground for this conclusion as follows:

> The sort of meaning that is basic to translation, and to the learning of one's own language, is necessarily empirical meaning. … Language is socially inculcated and controlled; the inculcation and control turn strictly on keying of sentences to shared stimulation. … Surely one has no choice but to be an empiricist so far as one's theory of linguistic meaning is concerned. (p. 81)

The idea is that since language is a tool for interpersonal communication the facts about meaning must be recoverable from intersubjective data. This conception of the ground of meaning facts had an enormous impact on Davidson. Davidson "thought it was terrific" and reported: "I sort of slowly put what I thought was

good in Quine with what I had found in Tarski. And that's where my general approach to the subject came from" (2004, p. 258).

Radical interpretation is similar to radical translation. For each, the evidence ultimately available consists in a speaker's dispositions to verbal behavior. But whereas the radical translator aims to produce a translation manual, the radical interpreter seeks to confirm an interpretive truth theory. And whereas the radical translator keys his translations to responses to patterns of stimulus at the sensory surfaces, the radical interpreter rather keys his interpretation to the speaker's responses to distal events—events in the shared environment.

Central to the radical interpreter's project is the confirmation of a Tarski-style axiomatic truth theory for the speaker's language. But this is not all that the interpreter aims to do. He must also fill in the picture of the speaker as a rational agent responding to his environment and others. Speaking is an activity embedded in a form of life appropriate for rational agents. As Davidson puts it at one point, "[a]ny attempt to understand verbal communication must view it in its natural setting as part of a larger enterprise" (2004, p. 151). This means that understanding what people mean by what they say must be fit into and be made coherent with a larger theory of them as rational beings.[7] Thus, in contrast to Quine's behaviorist approach, which eschewed any explicit appeal to psychological vocabulary, Davidson saw the framework of propositional attitude psychology and the explanation in its terms of behavior generally as the essential setting for understanding language.

Davidson characterizes the project of radical interpretation in terms of two questions. First, what could we know that would enable us to interpret another speaker? Second, how could we come to know it? A straightforward answer to the first question would seem to be an interpretive truth theory as characterized above. The answer to the second then would be a description of how a radical interpreter could come to

[7] Davidson's influential work in the philosophy of action (2001a) bears on the framework of rational agency invoked here.

confirm one for a speaker. Davidson does not answer the first question in this way, but rather suggests that if a theory has met certain empirical constraints, it is interpretive: "The present idea is that what Tarski assumes outright for each T-sentence can be indirectly elicited by a holistic constraint," namely, "that the totality of T-sentences should (in the sense described above [i.e. by way of the procedure of the radical interpreter]) optimally fit evidence about sentences held true by native speakers" (1973, p. 139).

This gives rise to a puzzle, however. Suppose that the answer to the first question (what we could know that would enable us to interpret another) is that a truth theory (of such and such a sort) has been confirmed by a radical interpreter. Then the answer to the second question (how could we confirm what we could know) should be a description of how to confirm *that a theory has been confirmed by a radical interpreter* (Lepore & Ludwig, 2005, pp. 151-166). But the answer to the second question was *supposed* to be a description of *radical interpretation itself*. What we get instead is a description of how to confirm that a radical interpreter has confirmed a truth theory. Something has gone wrong.

To avoid this problem, we can advert to another idea Davidson has identified as important, namely, that the theory be lawlike in the sense that it be projectable to instances that have not yet been observed, i.e., that it make the correct predictions for future and counterfactual utterances. Then the revised answer to the first question would be: the simplest lawlike theory that accommodates the behavioral evidence. The problem, however, is that we have seen already that being lawlike is *not* sufficient for interpretiveness, and it is not easy to see why being the *simplest* lawlike theory, if there is one, guarantees interpretiveness if lawlikeness does not.

Our problem is that, on the one hand, taking the proposal Davidson makes literally, given his questions, directs attention at the wrong thing, while retreating to the requirement that we confirm a lawlike truth theory clearly falls short of what is needed for interpretiveness. What is the solution to these difficulties? To have a clear target for the radical interpreter to aim at, we should answer the *first* question by citing something *uncontroversially* sufficient for interpretation, namely, an interpretive truth theory (and that it is interpretive

and so on, as above). We beg no questions by specifying that as the interpreter's aim, for to say that is his aim is not to say what the correct theory is for any given speaker. This has the added benefit that it gives us a clear standard by which to judge whether the interpreter's evidential base and constraints suffice for confirming something sufficient for interpretation.

To return to the radical interpreter's procedure, although the radical interpreter's evidential base is ultimately purely behavioral evidence, Davidson helps himself "at an intermediate stage" (1975, p. 161) to knowledge of a speaker's hold-true attitudes toward sentences. A speaker holds true a sentence *s* iff he believes *s* to be true. Davidson assumes that hold-true attitudes can be identified on the basis of more primitive behavioral evidence and that, by and large, for each of his beliefs a speaker holds true a sentence that expresses it. A speaker holds true a sentence *s* if *s* means that *p* and he believes that *p*. The point of focusing on hold-true attitudes then is that one can know someone holds a sentence true without knowing what it means or what belief it is based on. This helps to focus the question of how the radical interpreter is to marshal his evidence in order to assign meanings to sentences and detailed contents to attitudes. The ultimate aim of this is to illuminate the concept of meaning and related concepts, which Davidson treats as theoretical relative to the interpreter's evidence, by showing what the empirical content of a theory deploying them is. The empirical content is revealed in the implications any given theory has for what behavior should be expected assuming the theory is true. This shows what patterns in the data the theoretical concepts pick out, that is, how they organize the data into patterns intelligible in their terms.

If we can identify hold-true attitudes, then we can correlate them with what is going on in the speaker's environment. We will look for those hold-true attitudes which vary with variation in the environment and aim to identify lawlike correlations expressible in sentences of the form (L).

(L) *x* holds true φ at *t* iff *p*

How do we get from data in this form to an assignment of meaning to the sentence held true and content to the belief on the basis of which it is held true? If we knew either, we could solve for the other, but we don't start out with knowledge of either.

To break into the circle of meaning and belief we need to bring to bear a theoretical principle. Davidson proposed the Principle of Charity, according to which a speaker is largely rational and mostly right about his environment.[8] The Principle of Charity fixes belief to solve for meaning. If a speaker's beliefs about his environment are largely correct, then we can tentatively read the content of his beliefs off from the conditions correlated with the corresponding hold-true attitudes. The sentences held true then express those conditions. From correlations of the form (L) we can infer tentatively, where 'L' designates the subject's language, that (T) is a target theorem for an interpretive truth theory for L.

(T) For any speaker *x*, and time *t*, *s* is true in L at *t* for *x* iff *p*

Once the interpreter has identified target theorems of an interpretive truth theory, the interpreter formulates axioms that entail those theorems. These are used to make further predictions about behavior. Since people make mistakes, some of the beliefs that were initially treated as true may come to be treated as false, if that makes better overall sense of the speaker. In this way, the theory is adjusted until it achieves within the theoretical constraints an optimal fit with the evidence.

The Principle of Charity is the single most important theoretical upshot of taking the position of the radical interpreter as conceptually basic in understanding language. Davidson assumes that, as it is necessary for interpretation, being largely right about one's environment is constitutive of being a speaker. This entails that massive error in our empirical beliefs is incompatible with our possessing a language, and that as language-speakers our thoughts are relationally individuated: for with the same internal physical

[8] The idea and the label are inspired by Quine's Principle of Charity, "[A]ssertions startlingly false on the face of them are likely to turn on hidden differences of languages" (1960, p. 59). Davidson later distinguished the two elements mentioned here into the Principle of Correspondence and the Principle of Coherence (2001c, p. 211), the first of which moves toward Richard Grandy's the Principle of Humanity (Grandy, 1973).

states, a creature interpretable in a radically different environment will have radically different thoughts. Seeing the interpreter's standpoint as conceptually basic represents a complete reorientation of thinking about the relation of the mind to the world that is profoundly anti-Cartesian in the sense that it represents our epistemic and conceptual starting point as being, not the first-person point of view of introspection, but instead the third-person point of view of the interpreter of another.

7. Indeterminacy and the measurement analogy

A startling consequence of taking the radical interpreter's evidential position to be conceptually basic might be thought to call into question this fundamental assumption of the project. For after all the data is in, there will remain a range of incompatible interpretation theories which are empirically equivalent. One way this can happen is by making different choices about when to suppose someone is mistaken about something in the environment, as opposed to means something else by his words. A more fundamental source of the underdetermination of theory by evidence is the fact that it seems in principle possible to start with very different sets of (L) sentences. For if there is a correlation of a hold-true attitude with one condition in the environment, there will be many others. To use Quine's example, rabbits co-occur with undetached rabbits parts, the instantiation of rabbithood, and time slices of a rabbit. Beyond this, typically when a thing is causally responsible for another it is so only relative to background conditions. Our (L) sentences must be taken to hold only relative to those conditions—we do not *always* hold-true 'There is a rabbit' when there are rabbits around. But by focusing on something that is in the background relative to one (L) sentence, and bringing it to the foreground, while letting the previously correlated condition recede into the background, we can formulate a different (L) sentence. And if we do this systematically, we may arrive at very different interpretation theories for the speaker's language.

This is serious difficulty. From the point of view of the interpreter himself, the theories do not assign the same interpretations to object language sentences, because they represent them as being about different conditions in the environment. Prima facie, the conclusion should be that the radical interpreter's evidential

base is inadequate to confirm an interpretive truth theory and that it cannot exhaust the relevant evidence. This would undermine the view that the concepts of the theory of interpretation are theoretical concepts whose content is exhausted by the organization they impose on the behavioral data, and likewise the conclusions that follow from treating the Principle of Charity as constitutive of the interpreter's subject matter, since it could no longer be justified by holding it is necessary in order for radical interpretation to succeed.

However, Davidson held, following Quine's lead, that since the meaning facts are exhausted by the facts available to the radical interpreter, if different interpretation theories work equally well after all the evidence is in, they all capture the facts of the matter equally well. This is not underdetermination of theory by evidence, but *indeterminacy* of interpretation, in the sense that there is no determinate fact of the matter as to which of the range of empirically adequate theories is correct. Davidson aimed to render indeterminacy non-threatening. It is no more a problem, he maintained, than the fact that we can use either the Fahrenheit or Centigrade scales in keeping track of temperatures. We do not contradict each other when you say that it is 50 degrees Fahrenheit and I say that it is 10 degrees Celsius. We use relations among numbers to keep track of relations among temperatures by giving a physical interpretation to the use of numerals, for example, in terms of the height of a column of mercury. The numbers have more structure than what we use them to keep track of, so the choice of the mapping is only partly constrained by the phenomena. We must make some initial, arbitrary choices, such as to what temperature to assign 0 and what difference in temperature corresponds to the interval between 0 and 1. Relative to those choices, we can recover the empirical content of the use of a number to specify a temperature. So it is, Davidson claims, for different interpretation theories which are confirmable form the radical interpreter's standpoint. We use our sentences with their properties to keep track of other speakers' behavior. The semantic structure of our language is richer than the behavior. As an analogy, think of our practice of attributing attitudes to animals. Rover wags his tail upon hearing a car pull into the driveway. We keep track of his

behavior equally well by saying that he thinks his master is home, or that he thinks his provider is home. The behavioral evidence won't distinguish between these. Each attribution will work to track Rover's behavior provided we are systematic. This renders innocent the differences between schemes we might use in keeping track of Rover's behavior.[9] So also, the thought is, in the case of an interpreter of another speaker. We make certain arbitrary choices at the outset. Then we make further assignments of our words and sentences to those of the speaker we are interpreting in light of initial choices. As long as we keep track of the initial arbitrary choices, we can see how the different elaborations keep track of the same underlying phenomena.

It is far from clear that the measurement analogy is an adequate response to the difficulty. For it requires the interpreter to suppose that his own language has a richer semantic structure than that of the speaker he is interpreting. But the interpreter himself speaks a language. And the argument for underdetermination applies as well for any speaker of his language who attempts to interpret him from the standpoint of the radical interpreter. Since by hypothesis such a speaker has as rich a language, and since it must be possible for another to speak the interpreter's language, we have to conclude that the position of the radical interpreter is not after all adequate to confirm a interpretation theory for any language. For here the supposition required by the analogy is false.

It is natural to search for additional constraints. In later work, Davidson suggested that factoring in another speaker would help to narrow down choices. For we can, he says, identify the object of a speaker's thought with the common cause of a common response of the speaker and another with whom he is interacting (Davidson, 1991a, 1991b). They triangulate on an object, which is identified as what each is thinking about. It is doubtful that this gets us much traction, however. If two people sit in front of a television watching the news, for example, there will be many common causes of their common responses: events at

[9] Davidson held that animals do not have propositional attitudes, so this is not an illustration of the idea he would have endorsed himself.

the screen, in the television, in the signal from the TV station, in the news room, in satellites in orbit, and in distant trouble spots around the world. Another natural response would be to appeal to others being conspecifics and so very much like us in basic interests and in what they find salient. This would radically narrow down what to correlate their attitudes with. This way out, however, gives up on taking the standpoint of the radical interpreter, which takes the third person behavioral facts to determine the meaning facts, to be the fundamental standpoint from which to investigate meaning, for it allows that the first person stance on our own thoughts plays a fundamental role in interpretation.

Thus, in the end, it is unclear that the austere starting point of the radical interpreter puts one in a position to confirm something sufficient for interpreting another speaker's utterances, and to that extent it is unclear that the pure third person point of view suffices for understanding thought and language.

8. Davidson's Place in 20th Century Philosophy

Davidson sought to revolutionize the theory of meaning by effecting three reorientations in our thinking about it. The first is the introduction of the truth theory as the vehicle for the meaning theory, which aims to recover from the resources of the theory of reference all that we want from a compositional meaning theory. The second is the rejection of reductive analysis of 'is meaningful' in favor of a looser and more holistic form of conceptual illumination as represented by the application of an interpretation theory as a whole to the evidence as a whole. The third is the restriction of the evidence for the theory to what is available from the third person point of view without any presuppositions about meaning or the psychology of the speaker. Though still often misunderstood, Davidson's proposal that insight into meaning in natural languages can be obtained by reflection on how to construct and to confirm axiomatic truth theories for them has been hugely influential, both as a framework for doing natural language semantics and as a foil for critical discussion of foundational issues in the theory of meaning. Truth-theoretic semantics now is one of the paradigms for doing natural language semantics: its philosophical foundations and implications continue to attract debate; it has helped to clarify the distinction between investigations of logical form and lexical

analysis; and a great deal of detailed and fruitful work on natural language semantics has been undertaken within the framework, including important contributions by Davidson. Radical interpretation and the lessons Davidson drew from it have been more controversial, as it deals with a fundamental issue in thinking about the relation of thought and language to the word, namely, whether the concepts we use to describe these are properly thought of as deployed in the first instance from the third person point of view. Though there are difficulties in seeing how to make good on the idea, it would hard to overemphasize its importance.

Davidson's project can be seen, ironically, as a development of the broadly empiricist arc of analytic philosophy in the 20th century. Quine was the greatest influence on Davidson, as Carnap was on Quine. Carnap's outlook was structured by acceptance of the analytic–synthetic distinction and the view that the meaning of synthetic statements lies in their implications for sensory experience. This underwent two transformations in Quine. The first was his rejection of the analytic–synthetic distinction. The investigation of meaning then becomes a broadly empirical enterprise continuous with science and subject to considerations of fit with the rest of our empirical theory of the world. This gave rise to the second transformation, motivated by the view that language is a social art, which makes sensory experience appear an unsuitable basis for meaning, namely, the keying of meaning not to sensory experience but to stimulation of sensory surfaces. This represents a conservative modification of empiricist doctrine. Davidson takes over the third person stance from Quine, but the basis of meaning takes another step toward objectivity in being keyed to distal stimuli in the environment of speaker and interpreter. In this final step, the last vestiges of the traditional role of sensory experience in empiricism, as the basis of meaning and knowledge of the external world, are relinquished. Traditional empiricism, with its emphasis on the foundational role of experience in understanding meaning and knowledge, is turned on its head through a series of internal changes by which the third person point of view becomes conceptually, epistemically, and ontically basic. In this, Davidson completes the transformation of a fundamental philosophical view into

something so remote from its progenitor that its provenance can only be determined by tracing out the incremental steps by which it was accomplished.

References

Barwise, J. O. N., & Perry, J. (1981). Semantic Innocence and Uncompromising Situations. *Midwest Studies in Philosophy, 6*, 387-403.

Burge, T. (1992). Philosophy of Language and Mind: 1950-1990. *The Philosophical Review, 101*(1), 3-51.

Chihara, C. S. (1975). Davidson's extensional theory of meaning. *Philosophical Studies, 28*, 1-15.

Cummins, R. (2002). Truth and Meaning. In J. K. Campbell, M. O'Rourke & D. Shier (Eds.), *Meaning and Truth: Investigations in Philosophical Semantics* (pp. 175-193). New York: Seven Bridges Press.

Davidson, D. (1963). The Method of Extension and Intension. In A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap*. La Salle: Open Court.

Davidson, D. (1965). Theories of Meaning and Learnable Languages. In Y. Bar-Hillel (Ed.), *Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science.* (pp. 383-394). Amsterdam: North Holland Publishing Co. Reprinted in *Inquiries into Truth and Interpretation*.

Davidson, D. (1967). Truth and Meaning. *Synthese, 17*, 304-323. Reprinted in *Inquiries into Truth and Interpretation*.

Davidson, D. (1973). Radical Interpretation. *Dialectica, 27*, 314-328. Reprinted in *Inquiries into Truth and Interpretation*.

Davidson, D. (1975). Thought and Talk. In S. Guttenplan (Ed.), *Mind and Language*. Oxford: Oxford University Press. Reprinted in *Inquiries into Truth and Interpretation*.

Davidson, D. (1986). A Nice Derangement of Epitaphs. In E. Lepore (Ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Cambridge: Blackwell.

Davidson, D. (1991a). Epistemology Externalized. *Dialectica, 45*(2-3), 191-202. Reprinted in *Subjective, Intersubjective, Objective.*

Davidson, D. (1991b). Three Varieties of Knowledge. *Philosophy, 30*(Supp), 153-166. Reprinted in *Subjective, Intersubjective, Objective.*

Davidson, D. (2001a). *Essays on Actions and Events* (2nd ed.). Oxford: Clarendon.

Davidson, D. (2001b). *Inquiries into Truth and Interpretation* (2nd ed.). New York: Clarendon Press.

Davidson, D. (2001c). *Subjective, Intersubjective, Objective*. New York: Clarendon Press.

Davidson, D. (2004). *Problems of Rationality*. Oxford: Oxford University Press.

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

Foster, J. A. (1976). Meaning and Truth Theory. In G. Evans & J. McDowell (Eds.), *Truth and Meaning: Essays in Semantics* (pp. 1-32). Oxford: Clarendon Press.

Glock, H.-J. (2003). *Quine and Davidson on language, thought, and reality*. Cambridge: Cambridge University Press.

Grandy, R. (1973). Reference, Meaning, and Belief. *The Journal of Philosophy, 70*, 439-452.

Horwich, P. (2005). *Reflections on Meaning*. Oxford: Oxford University Press.

Katz, J. (1982). Common Sense in Semantics. *Notre Dame Journal of Formal Logic, 23*(2), 174-218.

Lepore, E., & Ludwig, K. (2005). *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Clarendon Press.

Lepore, E., & Ludwig, K. (2007). *Donald Davidson: Truth-theoretic Semantics*. New York: Oxford University Press.

Loar, B. (1976). Two Theories of Meaning. In G. Evans & J. McDowell (Eds.), *Truth and Meaning: Essays in Semantics* (pp. 138-161). Oxford: Clarendon Press.
Quine, W. V. O. (1960). *Word and Object*. Cambridge: MIT Press.
Quine, W. V. O. (1969). Epistemology Naturalized *Ontological Relativity and Other Essays* (pp. 69-90). New York: Columbia University Press.
Soames, S. (1992). Truth, Meaning, and Understanding. *Philosophical Studies, 65*(1-2), 17-35.
Soames, S. (2008). Truth and Meaning: In Perspective. *Midwest Studies in Philosophy, 32*, 1-19.
Stich, S. (1976). Davidson's Semantic Program. *Canadian Journal of Philosophy, 6*, 201-227.
Tarski, A. (1944). The Semantic Conception of Truth and the Foundations of Semantics. *Philosophy and Phenomenological Research, 4*, 341-376.
Tarski, A. (1983). The Concept of Truth in Formalized Languages *Logic, Semantics, Metamathematics* (2nd ed., pp. 506). Indianapolis: Hackett Publishing Company.