10

Truth in the Theory of Meaning

ERNIE LEPORE AND KIRK LUDWIG

1. Introduction

The publication of "Truth and Meaning" (TM) (2001c) in 1967 revolutionized work in the theory of meaning in three ways. First, it focused attention on the task how to understand compositionality in natural languages. Second, it advanced a powerful critique of the appeal to meanings, construed as entities, in the theory of meaning. Third, it proposed that the problem of compositionality be approached by constructing axiomatic truth theories for natural languages, modeled on the sort of truth theory that Alfred Tarski (1983) had shown how to construct for formal languages. The positive proposal, in particular, has been enormously influential. However, there has also been considerable controversy over exactly how to understand it.

There are two main interpretive positions. The first is the replacement theory, according to which Davidson aimed to replace the theory of meaning with a theory of truth conditions on the grounds that the concept of meaning is too confused to be an appropriate target for theorizing about language (Chihara 1975; Cummins 2002; Glock 2003: 142ff.; Katz 1982; Soames 1992, 2008; Stich 1976). A variant is that Davidson was engaged in a Carnapian explication of meaning aiming to capture certain features of the ordinary concept for certain theoretical purposes, without aiming to retain everything it involves (Ebbs 2012).[1] The second is the traditional pursuit theory, according to which Davidson aimed neither to abandon nor to reduce meaning, but to pursue the traditional project in the theory of meaning by a bit of indirection, and, in particular, by showing how putting appropriate empirical and formal constraints on a truth theory for a language yields theorems that provide interpretations of its sentences. In this chapter, we defend the traditional pursuit theory.[2]

There are three sources for the replacement theory. This first is puzzlement about how constructing a truth theory could be pursuing the traditional project. The second lies in certain passages in TM especially that suggest that he is abandoning the theory of meaning. The third is an appeal to the historical context, in which it is natural to

A Companion to Donald Davidson, First Edition. Edited by Ernie Lepore and Kirk Ludwig.

suppose that Davidson, who acknowledges Quine's influence (see Chapter 34 in this volume), was taking his lead from Quine (1960) in aiming to replace the ordinary notion with a new and more scientifically respectable concept.

In the following, we show, first, that the positive proposal can be seen as a straightforward pursuit of the traditional project, and, second, that the replacement theory fails to take into account the context of the passages which are its source in TM and the larger context of Davidson's work. In Section 2, we consider the immediate context of TM and its ostensible project: to spell out how one can provide a constructive account of the meanings of sentences in natural languages. In Section 3, we review Davidson's criticisms of then extant approaches to the theory of meaning and, with this as a background, make the case, in Section 4, that Davidson's proposal aims not to replace the theory of meaning but only to bypass difficulties that the direct approach encounters. In Section 5, we review passages that have been the primary source of the replacement theory and argue that the only the traditional pursuit theory makes good sense of them in their context. In Section 6, we show that Davidson's remarks about his project in later work, especially "Radical Interpretation" (RI) (2001d) and "Reply to Foster" (RF) (2001e), support this conclusion. In Section 7, we summarize.

2. Compositionality

In "Theories of Meaning and Learnable Languages" (TMLL) (2001b), published in 1965, Davidson writes:

> I propose . . . a necessary feature of a learnable language: it must be possible to give a constructive account of the meaning of sentences in the language. Such an account I call a theory of meaning for the language, . . . a theory of meaning that conflicts with this condition . . . cannot be a theory of a natural language. (Davidson 2001f: 3; all citations are to this volume)

The argument that a constructive account of the meanings of sentences in natural languages is possible rests on the fact we are finite beings and that natural languages have an infinity of nonsynonymous sentences. Understanding them must then rest on grasp of a finite number of semantical primitives and rules governing their combination (pp. 8–9). Davidson characterizes the project abstractly in the following passage:

> we are entitled to consider in advance of empirical study . . . how we shall describe the skill or ability of a person who has learned to speak a language. . . . [A natural condition on this] is that . . . our theory should equip us to say, for an arbitrary sentence, what a speaker of the language means by that sentence (or takes it to mean). Guided by an adequate theory, we see how the actions and dispositions of speakers induce on the sentences of the language a semantic structure. (p. 8)

We call such a theory for a particular language a meaning theory. We use "theory of meaning" for the project of explaining meaning in general. While Davidson goes on to criticize, with this requirement in mind, a number of proposals about the meanings of constructions in natural languages (pp. 9–15), the importance of these passages here lies in the context they provide for interpreting TM. For clearly, this project focuses on understanding *meaning* in natural language, and Davidson is not rejecting it but arguing that it should take a particular form. That this is the *same project* as that of TM is established by three considerations. First, these papers were written at roughly the same time. Second, in explaining his project at the beginning of TM, Davidson refers back to the conclusion he reaches in TMLL.

> It is conceded by most [theorists] that a satisfactory theory of meaning must give an account of how the meanings of sentences depend upon the meanings of words. Unless such an account could be supplied for a particular language . . . there would be no explaining the fact that we can learn the language: no explaining the fact that, on mastering a finite vocabulary and a finitely stated set of rules, we are prepared to produce and understand any of a potential infinitude of sentences. I do not dispute these vague claims, in which I sense more than a kernel of truth [and here Davidson cites TMLL in a footnote]. Instead I want to ask what it is for a theory to give an account of the kind adumbrated. (p. 17)

Third, he suggests already in TMLL that a truth theory can provide an appropriate vehicle for carrying out the project (p. 8). Thus, it is clear that these two papers are parts of a single project. The first argues for a condition on an adequate meaning theory, and the second takes up the project of saying how it could be carried out. At the end of TM, he writes,

> In this paper I have assumed that the speakers of a language can effectively determine the meaning or meanings of an arbitrary expression (if it has a meaning), and that it is *a central task of a theory of meaning to show how this is possible.* I have argued that *a characterization of a truth predicate describes the required kind of structure, and provides a clear and testable criterion of an adequate semantics for a natural language.* No doubt there are other reasonable demands that may be put on a theory of meaning. But *a theory that does no more than define truth for a language comes far closer to constituting a complete theory of meaning than superficial analysis might suggest; so, at least, I have urged.* (p. 35; emphasis added)

That is to say, the project was to show how to give a compositional meaning theory, and the argument of the paper aimed to show that a truth theory could go a long way to doing *that job.* We will see how in Section 4.

3. Criticism of Traditional Approaches

Many interpreters treat TM as if it began five pages into the essay with the positive proposal. But the initial critical discussion of attempts to provide a meaning theory by assigning entities to expressions is essential for understanding that proposal for at least three reasons. First, it is essential for understanding what Davidson found inadequate in prior approaches, namely, the futility of quantifying over expression meanings. Second, it lays the groundwork for the positive proposal by working out a characterization of the project that does not build into its statement the requirement that things called meanings be associated with every sentence. Third, it provides, in the context of a discussion of a reference theory for a fragment of a language, a simplified model for the full proposal.

A plausible constraint on a constructive theory is that it entails, from axioms about semantical primitives, all true sentences of the form: s means that p (or a generalization to allow for context sensitivity; see (Lepore and Ludwig 2007: chapter 1, sections 6–8). An initially attractive approach is to take every expression to refer to its meaning, and to take meanings of complex expressions to be constructed out of the meanings of their parts. An initial problem is how this could distinguish between a sentence and a list (p. 17): What distinguishes "Theatetus flies" from "the meaning of *Theatetus* the meaning of *flies*"? The moral Davidson draws is that we need a rule that tells us how to go from terms in different categories (referring term and predicate) to one in a third (a sentence), which is evaluable as true or false (p. 18). Once the need for a rule is recognized, he argues, it becomes clear that assigning an entity to every expression is not needed.

He illustrates the point with a simple reference theory (pp. 18–19). To bring out the full force of the example, we modify it by distinguishing the language of the theory (the metalanguage) from the language it is about (the object language). Consider a reference theory (in English) for a fragment of French consisting of the singular terms "Marie" and "Annette" and the functor "la mére de," which, when concatenated with a singular term, yields a singular term. We specify that "Marie" refers to Marie and that "Annette" refers to Annette. To show the insufficiency of just assigning entities to expressions for understanding (even for a reference theory), suppose that "la mere de" refers to a function, and we heed the advice to introduce a rule specifying referents of complex singular terms:

> the concatenation of 'la mére de' with a singular term t refers to the value of the function referred to by 'la mére de' given the referent of t as argument.

This gives the referent *in a sense*. But it gives us no insight into it. Even here it is clear we want to *understand* what the term refers to. The obvious response is to add that *the function takes individuals to their mothers*. But now the function drops out as unnecessary, for we can give instead the rule:

> the concatenation of 'la mére de' with a singular term t refers to the mother of the referent of t.

If the goal of the theory can be accomplished without assigning an entity to every expression, then the success condition should be stated so as not to require this. Davidson suggests (when object and metalanguage are identical) we require that

> a theory entails every sentence of the form 't refers to x' where 't' is replaced by a structural description of a singular term and 'x' is replaced by that term itself (p. 18).

When, as in our example, object and metalanguage are distinct, we require (what the syntactic condition in the former case secures) that

> a theory entails every sentence of the form 't refers to x' where 'x' is replaced by a term that translates t.

implified model for

rom axioms about
p (or a generaliza-
hapter 1, sections
fer to its meaning,
of the meanings of
a sentence and a
g of *Theatetus* the
e that tells us how
e) to one in a third
need for a rule is
every expression is

). To bring out the
age of the theory
guage). Consider a
the singular terms
concatenated with
s to Marie and that
signing entities to
e that "la mere de"
cifying referents of

e of the function

ven here it is clear
e is to add that *the*
out as unnecessary,

er of the referent

an entity to every
ot to require this.
e require that

ed by a structural
18).

require (what the

eplaced by a term

We now make an observation Davidson does not, but must have had in mind, that helps to illuminate how he thought a truth theory, ostensibly falling "comfortably within the theory of reference" (p. 23), could discharge the duties of a meaning theory. The key lies in the requirement that the simple reference theory use metalanguage terms that translate object language terms in giving their referents. This explains why it gives genuine *insight into* referents of object language terms. But this also means that, knowing this, we can read off from the theorems what the object language singular terms *mean*. For if we know that "La mére de Marie" refers to the mother of Marie, and we know that in stating that we have given the referent using a term that translates the object language term, we can see also that "La mére de Marie" means *the mother of Marie*. We have then squeezed out of a reference theory an interpretation of each singular term in the language fragment. We have achieved the goal through a bit of indirection. This is Davidson's key idea, and we will see how it is pressed into service in the positive proposal.

Before turning to the positive proposal, Davidson rejects two other suggestions and specifies a desideratum on an adequate account, which prefigures the condition of success it turns out a suitable truth theory can meet.

The first proposal is the suggestion that predicates be treated on analogy with the functor "la mere de" and sentences as referring terms, with the hope of reduplicating the success of the reference theory for the full language. Davidson rejected this with an argument, famously dubbed the slingshot (Barwise and Perry 1981), designed to show that if sentences refer to anything, sentences alike in truth value refer to the same things, "an intolerable result" (p. 19). The argument assumes that logically equivalent singular terms corefer and that replacing a singular term in another does not affect reference of the containing term. Then, any sentence "R" is logically equivalent to "{x: x = x & R} = {x: = x}"; substituting "{x: x = x & S}" for the first singular term does not affect reference if "S" and "R" are alike in truth value; the result is logically equivalent to "S." So, any two sentences alike in truth value refer to the same thing. However, this presupposes that sentential logical equivalence, sameness of *truth value* under all reinterpretations of nonlogical terms, suffices for logical equivalence of singular terms, sameness of *reference* under all reinterpretations of nonlogical terms, which begs the question. Despite this, there is remarkably little to be said for sentences being referring terms, and so we may, in any case, set the proposal aside as misguided.

Dropping sentential reference, one might simply appeal to a function from meanings of parts to meanings of complexes. Yet, it is no help just to say: the meaning of "Theatetus flies" is the value of the meaning of "flies" given the meaning of "Theatetus" as argument, for this (again) does not help us understand what "Theatetus flies" means. As Davidson sums it up,

> What analogy 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. Clearly some more articulate way of referring to meanings than any we have seen is essential if these criteria are to be met. (p. 20)

Davidson has no objection to a theory taking this form except that we have no way of picking out meanings of sentences that enables us to understand them. He concludes:

> Paradoxically, the one thing *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. My objection to *meanings* in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use. (pp. 20–21; emphasis added)

The charge is that *meanings*, construed as entities, do not advance the project of *giving the meaning* of sentences in the sense of enabling us to interpret them. There is no suggestion here that we abandon the *theory of meaning* as opposed to giving up the fruitless appeal to *meanings*.

The second proposal is to add a dictionary to a recursive syntax. But "Hopes will be dashed . . . if semantics is to comprise a theory of meaning in our sense, for knowledge of the structural characteristics that make for meaningfulness in a sentence, plus knowledge of the meanings of the ultimate parts, does not add up to knowledge of what a sentence means" (p. 21), at least in the sense of being able to understand it. The trouble is that the syntax, which specifies meaningful strings, does not give us any rules for interpreting complexes on the basis of their parts. Davidson goes on to say that while "there is agreement that it is the central task of semantics to give the semantic interpretation (the meaning) of every sentence in the language, nowhere in the linguistic literature will one find, so far as I know, a straightforward account of how a theory performs this task, or how to tell when it has been accomplished" (p. 21). This again shows that Davidson is interested in a theory that gives semantic interpretations of sentences from a finite basis, in the sense of putting us in a position to interpret them. The problem he addresses is how to do that.

4. The Positive Proposal

The immediate preamble to the positive proposal is often cited as the place Davidson abandons the theory of meaning. But a careful look at this and the following paragraphs (pp. 22–23) show that something more subtle and interesting is going on:

> [the suggestion was that] an adequate theory of meaning must entail *all* sentences of the form '*s* means *m*'. But now, having found no more help in meanings of sentences than in meanings of words, let us ask whether we can get rid of the troublesome singular terms supposed to replace '*m*' in '*s* means *m*' and to refer to meanings. In a way, nothing could be easier: just write '*s* means that *p*', and imagine '*p*' replaced by a sentence. Sentences, as we have seen, cannot name meanings, and sentences with 'that' prefixed are not names at all. . . . It looks as though we are in trouble on another count, however, for it is reasonable to expect 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)

that we have no
lerstand them. He

els of a theory of
n-trivially give the
: in the theory of
: obscure, but that

:he project of *giving*
m. There is no sug-
/ing up the fruitless

But "Hopes will be
ense, for knowledge
n a sentence, plus
knowledge of what
understand it. The
iot give us any rules
on to say that while
the semantic inter-
ere in the linguistic
nt of how a theory
' (p. 21). This again
c interpretations of
n to interpret them.

the place Davidson
and the following
and interesting is

all sentences of the
f sentences than in
ome singular terms
way, nothing could
tence. Sentences, as
ixed are not names
ever, for it is reason-
-extensional 'means
h, the problems our

Two proposals are dismissed, but they are proposals for how to achieve the goal of giving a compositional meaning theory. The first appeals to meanings as entities, but these have proved no help. The second, a theory with theorems of the form "s means that p," avoids this difficulty but raises another. Without a referring term after "means," we must formulate a logic for replacements in such contexts sensitive to what terms mean, for *complex* as well as simple terms. This requires a prior analysis of semantic structure, however, of just the sort the theory is to provide. How then to achieve our goal? It is in this light that the next passage should be understood (roman numerals added).

> (i) The only way I know to deal with this difficulty is simple, and radical. (ii) Anxiety that we are enmeshed in the intensional springs from using the words "means that" as filling between description of sentence and sentence, but it may be that the success of our venture depends not on the filling but on what it fills. (iii) The theory will have done its work if it provides, for every sentence *s* in the language under study, a matching sentence (to replace "*p*") that, in some way yet to be made clear, "gives the meaning" of *s*. (iv) 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. (v) 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. (vi) The plausible result is
>
> (T) *s* is *T* if and only if *p*
>
> (vii) 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.

The difficulty in (i) is just that of avoiding the dilemma sketched in the preamble. (ii) remarks that perhaps the crucial thing is not so much the use of "means that" but the matching thereby achieved of a mentioned object language sentence with a used metalanguage sentence that (we know) interprets it. For (iii), we want a theory that "gives the meaning" of the sentence in the sense of enabling us to understand it (the parenthetical "in a sense to be made clear" has to do with eliminating the lingering use of "meaning" as a count noun). An obvious candidate (iv), if the metalanguage contains the object language, is to match the mentioned sentence with the sentence itself in use, and otherwise a *translation* of it (recall the reference theory). To avoid the "springs of the intensional," we should then (v) replace "means that" with a sentential connective ("if and only if" being the obvious candidate), and apply a predicate to the mentioned sentence so that we have a sentence on each side of the connective. (vi) The result is (T). A theory that met this constraint (and was otherwise formally correct) would, as Davidson immediately notes (and had in mind all along), satisfy Tarski's Convention T: "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" (p. 23).

Convention T requires an adequate theory (consisting of base axioms for referring terms and predicates and recursive axioms for logical connectives, quantifiers, etc.) to entail all theorems of the form (T) in which "s" is replaced by a description of an object language sentences as composed out of its significant parts and "p" by a metalanguage sentence translating it. An axiomatic theory of a predicate "is true" meeting this condition has all and only the true sentences of the language in its extension (setting aside the semantic paradoxes and context sensitivity). The right-hand sides of its canonical theorems (those of form (T) whose proofs draw minimally on the content of the axioms) would use sentences that translated the sentence mentioned on the left. Having such a theory and knowing that we did (and a proof procedure for canonical theorems) would enable us to "give the meaning" of each object language sentence in the sense of being in a position to interpret it. In fact, since the relation that Convention T requires between s and "p" in relevant instances of (T) is exactly that required between them in (M) "s means that p," we could infer (M) from (T), as Davidson notes in "Semantics for Natural Languages" (1970; first read in 1968 – the year after TM was published):

> A theory of truth entails, for each sentence *s*, a statement of the form '*s* is true if and only if *p*' where in the simplest case '*p*' is replaced by *s*. Since the words 'is true if and only if' are invariant, we may interpret them if we please as meaning 'means that'. So construed, a sample might then read '"Socrates is wise" means that Socrates is wise'. (p. 60)

We have thus found a promising approach to avoiding the two horns of the dilemma in the preamble.[3]

5. Problematic Passages?

Davidson's announced project is providing a compositional meaning theory for a natural language. He develops a dilemma for the project, and then he urges a way around it by a bit of indirection, namely, by constructing an axiomatic truth theory for the language that meets Convention T. Seeing how this works to achieve the aims of the announced project undermines the first of the motivations for the replacement theory, and the argument from historical context cannot stand on its own. This leaves the passages alluded to earlier which follow the transitional passage, which have been a rich source for the replacement theory.

First, though, in the transitional passage itself [vii], Davidson requires that "without appeal to any (further) semantical notions," a theory of meaning put enough restrictions on the predicate "is T" to satisfy Convention T. What could this signify except the desire to eschew the concept of meaning? But this makes no sense of our requiring something of a *theory of meaning*. Rather, at this point, we suggest, Davidson had in effect shifted from focusing on the initial project of just giving a compositional meaning theory to the extended project of illuminating more generally what it is for words to mean what they do (see Davidson 2001a: xiii). He hoped that once we adjusted a truth theory to handle context-sensitive constructions and treated it as an empirical theory of a speaker (as elaborated in "RI" – see Chapters 13–14 in this volume), merely getting a workable theory would guarantee that it had met Convention T. This would promise

xioms for referring
quantifiers, etc.) to
ription of an object
by a metalanguage
meeting this condi-
ısion (setting aside
les of its canonical
ıtent of the axioms)
left. Having such a
al theorems) would
ı the sense of being
T requires between
een them in (M) "s
nantics for Natural
ished):

is true if and only
rue if and only if'
ıat'. So construed,
se'. (p. 60)

rns of the dilemma

ıning theory for a
en he urges a way
ıtic truth theory for
o achieve the aims
for the replacement
its own. This leaves
e, which have been

ıuires that "without
put enough restric-
is signify except the
se of our requiring
st, Davidson had in
ıpositional meaning
at it is for words to
we adjusted a truth
an empirical theory
ıme), merely getting
This would promise

illumination of meaning in terms of more basic concepts, and so not just illuminate how we understand complexes on the basis of their parts but also semantical primitives. Without this additional ambition that enters at this point, we could instead simply require that the axioms of the truth theory themselves meet an analog of Convention T, requiring them to use metalanguage terms that interpret object language terms for which they give reference, truth and satisfaction conditions (Lepore and Ludwig 2005: 109, chapter 4, section 4, 2007: chapters 1–3, esp. chapter 3, section 4). This turns out to be important for understanding what follows.

One of the chief sources of the view that Davidson has abandoned the theory of meaning in favor of a different project lies in the passage following the transitional paragraphs (labels added):

> [a] There is no need to suppress, of course, the obvious connection between a definition of truth of the kind Tarski has shown how to construct, and the concept of meaning. [b] It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. [c] To know the semantic concept of truth for a language is to know what it is for a sentence—any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language. [d] This at any rate is my excuse for a feature of the present discussion that is apt to shock old hands; my freewheeling use of the word 'meaning', for what I call a theory of meaning has after all turned out to make no use of meanings, whether of sentences or of words. [e] Indeed, since a Tarski-type truth definition supplies all we have asked so far of a theory of meaning, it is clear that such a theory falls comfortably within what Quine terms the 'theory of reference' as distinguished from what he terms the 'theory of meaning'. [f] So much to the good for what I call a theory of meaning, and so much, perhaps, against my so calling it. (p. 24)

This should be read in the light of what precedes it. [b] and [c] have been taken to suggest that Davidson abandons meaning for truth conditions: But given that we have in mind truth conditions stated in canonical theorems of a truth theory meeting Convention T (i.e., "the semantic concept of truth"), it is clear how this "gives the meaning" of the object language sentence. [d] likewise has been cited as evidence that Davidson throws out the theory of meaning, using the word but discarding its traditional use: But it is clear that it is meanings in the plural, that is, as entities, that he rejects, as is clear in the critical discussion. [e] and [f] should be interpreted in context as well: The theory of truth itself does not employ concepts other than those drawn from the theory of reference – it is only in light of knowledge that it satisfies Convention T that it puts us in a position to interpret object language sentences, and the theory does not state that it satisfies Convention T. Read in context, everything falls into place.

Davidson does make a mistake, which he notes later (Davidson 2001d: 138–139, 2001e: 171–173). He fails to distinguish between the knowledge we have about a suitable truth theory (that it meets constraints sufficing for it to satisfy Convention T), and what the truth theory itself states. He puts it this way:

> My mistake was not . . . to suppose that any theory that correctly gave truth conditions would serve for interpretation; my mistake was to overlook the fact that someone might know a sufficiently unique theory without knowing that it was sufficiently unique. The

> distinction was easy for me to neglect because I imagined the theory to be known by someone who had constructed it from evidence, and such a person could not fail to realize that his theory satisfied the constraints. (Davidson 2001e: 173)

Another passage that has fueled the replacement theory follows a paragraph in which Davidson says that the truth theory is an empirical theory, the empirical power of which "depends on success in recovering the structure of a very complicated ability—the ability to speak and understand a language" (p. 25). The striking thing Davidson says is

> [We ought not to be conned] into thinking a theory any more correct that entails '"Snow is white" is true if and only if snow is white' than one that entails instead:
>
> (S) 'Snow is white' is true if and only if grass is green.
>
> Provided . . . we are as sure of the truth of (S) as we are of that of its more celebrated predecessor. (pp. 25–26)

Further,

> The grotesqueness of (S) is in itself nothing against a theory of which it is a consequence, provided the theory gives the correct results for every sentence (on the basis of its structure, there being no other way). It is not easy to see how (S) could be party to such an enterprise, but if it were—if, that is, (S) followed from a characterization of the predicate 'is true' that led to the invariable pairing of truths with truths and falsehoods with falsehoods—then there would not, I think, be anything essential to the idea of meaning that remained to be captured. (p. 26)

Yet it seems clear that there would be much that was not captured about the meaning of "snow is white" in using "grass is green" to interpret it! How then can one construe what Davidson is saying sensibly without taking him to be simply abandoning the traditional project as intractable?

A footnote added in 1982 provides an essential clue:

> Critics have often failed to notice the essential proviso mentioned in this paragraph. The point is that (S) could not belong to any reasonably simple theory that also gave the right truth conditions for 'That is snow' and 'This is white'. (See the discussion of indexical expressions below.) (p. 26, note 10)

This shows that he does not think (S) could belong to an adequate theory. For an *empirical* theory, assigning truth conditions *relative to context* has to capture the "ability to speak and understand a language." It must then assign correct truth conditions to demonstrative sentences. A theory using "is green" to say how "is white" is used would incorrectly entail that uses of "That is white" would be true iff what was demonstrated was green. So the point is to express the view that a theory empirically adequate to a language actually spoken would also be adequate for interpretation, not to replace a thick notion of meaning with a newly contrived thin notion of equivalence in truth value (something Davidson clearly rejects in the slingshot).

The next paragraph is puzzling as well, but in a footnote added in 1982, Davidson says that it is simply confused.

> It would be ill advised for someone who had any doubts about the colour of snow or grass to accept a theory that yielded (S), even if his doubts were of equal degree, unless he thought the colour of the one was tied to the colour of the other. Omniscience can obviously afford more bizarre theories of meaning than ignorance; but then, omniscience has less need of communication. (pp. 26–27)

The footnote adds, however,

> This paragraph is confused. What it should say is that sentences of the theory are empirical generalizations about speakers, and so must not only be true but also lawlike. (S) presumably is not a law, since it does not support appropriate counterfactuals. It's also important that the evidence for accepting the (time and speaker relativized) truth conditions for 'That is snow' is based on the causal connection between a speaker's assent to the sentence and the demonstrative presentation of snow. For further discussion see Essay 12 ["Reply to Foster"]. (p. 26, n. 11)

Davidson's considered view is that the trouble with (S) is that it is not a law and so does not track the dispositions that guide our use of words, as is shown by the false predictions a theory generating it makes about demonstrative sentences. The point again is that a truth theory empirically confirmed for a speaker will *in fact* meet Convention T.

Davidson takes up this project later in RI, in which he explains the methodology as follows:

> In philosophy we are used to definitions, analyses, reductions. Typically these are intended to carry us from concepts better understood, or clear, or more basic epistemologically or ontologically, to others we want to understand. . . . I have proposed a looser relation between concepts to be illuminated and the relatively more basic. At the centre stands a formal theory, a theory of truth, which imposes a complex structure on sentences containing the primitive notions of truth and satisfaction. These notions are given application by the form of the theory and the nature of the evidence. The result is a partially interpreted theory. The advantage of the method lies not in its free-style appeal to the notion of evidential support but in the idea of a powerful theory interpreted at the most advantageous point. This allows us to reconcile the need for a semantically articulated structure with a theory testable only at the sentential level. The more subtle gain is that very thin evidence in support of each of a potential infinity of points can yield rich results, even with respect to the points. By knowing only the conditions under which speakers hold sentences true, we can come out, given a satisfactory theory, with an interpretation of each sentence. It remains to make good on this last claim. The theory itself at best gives truth conditions. What we need to show is that if such a theory satisfies the constraints we have specified, it may be used to yield interpretations. (pp. 137–138)

The goal is to put nonsemantic constraints on a truth theory to yield interpretive truth conditions. This is a broadly empiricist approach to illuminating the concept of meaning, but it contains no commitment to any simple translation of talk of meaning into talk of patterns of behavior, or to any other concepts. If the approach is right, then truth and meaning are intimately connected, but neither is reducible to the other. (See Chapters 13–14 in this volume and (Lepore and Ludwig 2005, esp. chapters 11–16) for a fuller discussion of this aspect of the project.)

6. Later Work

Davidson's development of his project confirms the reading we have given of it in TM. In this section, we concentrate on RI and RF.

A problem with the suggestion that a correct truth theory is adequate for interpretation is that given one theory, we can generate another that is true iff it is by adding a true conjunct to the application conditions in any axiom for any predicate. If we start with [A1], we can find an extensionally equivalent theory by replacing it with [A2]:

[A1] "Snow" applies to something iff it is snow.
[A2] "Snow" applies to something iff it is snow and the earth moves.

But while [A1] could be party to a theory that met Convention T, [A2] could not. In RF (delivered in 1973 but published in 1976), Davidson remarks that in RI, he criticized his earlier attempts to say what the relation was between a truth theory and a meaning theory and "tried to do better" (p. 171). One of the problems was the extensionality problem just mentioned. This alone casts doubt on the replacement theory, for if Davidson were simply replacing meaning with truth, there would be no problem in the first place.

> RI begins with two questions:
>
> Kurt utters the words 'Es regnet' and under the right conditions we know that he has said that it is raining. Having identified his utterance as intentional and linguistic, we are able to go on and to interpret his words: we can say what his words, on that occasion, meant. What could we know that would enable us to do this? How could we come to know it? (p. 125)

It is clear then that the project of the paper is focused on what knowledge would suffice to interpret a speaker in the sense of putting us in a position to say "what his words, on that occasion, meant," and for assertion this is saying that his utterance meant that p, for some "p":

> What knowledge would serve for interpretation? A short answer would be, knowledge of what each meaningful expression means. In German, those words Kurt spoke mean that it is raining and Kurt was speaking German. So in uttering the words 'Es regnet', Kurt said that it was raining. (p. 126)

The trouble with this answer is not that it traffics in meaning, but that it does not say "what it is to know what an expression means" (p. 126). The appeal to assigning "to each meaningful expression . . . an entity, its meaning" is dismissed as "very little help" and as an expedient that "at best hypostatizes the problem" (p. 126), echoing the

oach is right, then
e to the other. (See
. chapters 11–16)

e given of it in TM.

uate for interpreta-
iff it is by adding a
edicate. If we start
ing it with [A2]:

ves.

2] could not. In RF
RI, he criticized his
ory and a meaning
the extensionality
eory, for if Davidson
m in the first place.

w that he has said
guistic, we are able
t occasion, meant.
come to know it?

vledge would suffice
ay "what his words,
tterance meant that

d be, knowledge of
rt spoke mean that
's regnet', Kurt said

that it does not say
eal to assigning "to
issed as "very little
p. 126), echoing the

criticism of TM. On developing the problem, it is clear that the "interpreter must be able to understand any of the infinity of sentences the speaker might utter" (p. 127), and so in explaining what enables him to do it, "we must put it in finite form" (p. 128). So we want a compositional meaning theory—the same project that TM opens with. In view of the inutility of meanings, we should "describe what is wanted . . . without apparent reference to meanings or interpretations: someone who knows the theory can interpret the utterance to which the theory applies" (p. 128). "A satisfactory theory . . . will [therefore] reveal significant semantic structure" (p. 130). If we had an interpretation theory for our language, and a translation theory for another language, we could provide interpretations of its sentences. But the reference to the home language is "an unneeded intermediary between interpretation and alien idiom." We should apply an interpretation theory directly to the foreign language: "what is left is a structurally revealing theory of interpretation for the object language. . . . We have such theories," Davidson suggests, "in theories of truth of the kind Tarski first showed how to give" (p. 130).

Thus, in RI, Davidson's proposal is that we would have an interpretation theory for a language that enables us to say what utterances in it mean, in the sense of enabling us to understand them, if we can construct for it a truth theory we know to satisfy Convention T. This is his answer to the first of the two questions posed at the outset of the paper. It is clearly continuous with the project of TM, in which he treats the truth theory as an empirical theory ultimately justified from the standpoint of radical interpretation (p. 27).

Davidson goes on to describe a procedure for confirming a truth theory from evidence that does not presuppose knowledge of meanings or detailed contents of attitudes. He asks whether a theory satisfying the constraints would enable us to interpret utterances of the language. Davidson is clear in RI, in contrast to TM, that just knowing the truth theory is not enough, for (p. 138) "a T-sentence does not give the meaning of the sentence it concerns: the T-sentences [sic] does fix the truth value relative to certain conditions, but it does not say the object language sentence is true because the conditions hold" (because the conditions are that p and it means that p).

> Yet if truth values were all that mattered, the T-sentence for 'Snow is white' could as well say that it is true if and only if grass is green or 2 + 2 = 4 as say that it is true if and only if snow is white. (p. 138)

Adding "the canonical proof of a T-sentence" does not help, for anomalous ones have proofs as well. "If we knew that a T-sentence satisfied Tarski's Convention T," Davidson notes, "we would know that it was true, and we could use it to interpret a sentence because we would know that the right branch of the biconditional translated the sentence to be interpreted" (p. 139). But this has to be confirmed in RI, so we have to say how we could come to know this. However,

> What we have been overlooking . . . is that we have supplied an alternative criterion: this criterion is that the totality of T-sentences should (in the sense described above) optimally fit evidence about sentences held true by native speakers. The present idea is that what Tarski assumed outright for each T-sentence can be indirectly elicited by a holistic constraint. If that constraint is adequate, each T-sentence will in fact yield an acceptable interpretation. (p. 139)

It is satisfaction of those constraints that is to ensure the theory meets Convention T and solve the extensionality problem. Knowledge that the theory meets the constraints plus knowledge of a canonical proof procedure puts us in a position to interpret any utterance of the language. This again shows Davidson is not abandoning the theory of meaning as hopeless, but arguing that it can be pursued in a deeply illuminating way by considering what empirical constraints on a truth theory would suffice for it to meet Convention T.

RF is Davidson's most explicit discussion of the goals and structure of his project, and he reiterates the position we have just outlined. He says, echoing the last paragraph of RI:

> Since Tarski was interested in defining truth . . . he could take the concept of translation for granted. But in *radical* interpretation, this is just what cannot be assumed. So I have proposed instead some empirical constraints on accepting a theory of truth that can be stated without appeal to such concepts as those of meaning, translation, or synonymy, though not without a certain understanding of the notion of truth. By a course of reasoning, *I have tried to show that if the constraints are met by a theory, the T-sentences that flow from that theory will in fact have translations of s replacing 'p'*. (p. 172; emphasis added).

The challenge Foster raises is not that the constraints Davidson gives are inadequate to respond to the extensionality problem (p. 173), but rather two nested problems, one of which he thinks can be met only at the cost of raising the other. The first is that Davidson treats knowledge of what the truth theory states as sufficient for interpretation, but it is not. The second problem is that what has to be added – knowledge that a truth theory meeting the constraints *states that* s is true iff p – is not available to Davidson because it draws on intensional notions. On the first, Davidson flatly denies that he ever thought that knowledge of what the truth theory states was sufficient (p. 174): "So far as I know, I never held the view . . . which leaves unconnected the knowledge of what a theory of truth states and the knowledge that the truth theory is T-theoretical" (i.e., satisfies Convention T). "The interpreter does, indeed, know that his knowledge consists in what is stated by a T-theory, a T-theory that is translational (satisfies Convention T)" (p. 175).

> Someone who can interpret English knows . . . that an utterance of the sentence 'Snow is white' is true if and only if snow is white; he knows in addition that this fact is entailed by a translational theory—that it is not an accidental fact about that English sentence, but a fact that interprets the sentence. Once the point of putting things this way is clear, I see no harm in rephrasing what the interpreter knows in this case in a more familiar vein: he knows that 'Snow is white' in English *means that* snow is white. (p. 175)

This is exactly the view that we attributed to him in TM. To the second charge, Davidson denies having the goal that Foster "foists" on him. He says that he does not "believe it is possible to reduce these notions [language, meaning, belief, and intention] to anything more scientific or behavioristic." Further, "What I have tried to do is to give an account of meaning (interpretation) that makes no essential use of unexplained linguistic concepts. . . . It will ruin no plan of mine if in saying what an interpreter knows it is necessary to use a so-called intensional notion—one that consorts with belief and intention and the like" (p. 176). The restriction Davidson has in mind is on the constraints a truth theory has to meet to satisfy Convention T, but not on specifying what the interpreter knows, and, in particular, it does not rule out his knowing that a theory that meets the constraints entails a certain thing (p. 178).

7. Conclusion

Davidson's project is not to reduce meaning to truth conditions or to replace the theory of meaning with a successor project more suitable to scientific progress, but a pursuit of a theory of meaning by a bit of clever indirection. This is not an esoteric doctrine of Davidson's. He announces it in TM and he explains it in the introduction to *Inquiries into Truth and Interpretation* (Davidson 2001a):

> What is it for words to mean what they do? In the essays collected here I explore the idea that we would have an answer to this question if we knew how to construct a theory satisfying two demands: it would provide an interpretation of all utterances, actual and potential, of a speaker or group of speakers; and it would be verifiable without knowledge of the detailed propositional attitudes of the speaker. The first condition acknowledges the holistic nature of linguistic understanding. The second condition aims to prevent smuggling into the foundations of the theory concepts too closely allied to the concept of meaning. A theory that does not satisfy both conditions cannot be said to answer our opening question in a philosophically instructive way. (p. xiii)

The project of explaining what it is for words to mean what they do is carried out by reflecting on how to construct and confirm from the standpoint of an interpreter a theory for particular languages without presupposing knowledge of what the theory is about. Davidson's suggestion is that a truth theory (known to satisfy Convention T or an analog for natural languages) would satisfy the first demand; he aimed to show how it could be verified from the standpoint of the radical interpreter to satisfy the second. That this is Davidson's project has nevertheless escaped many commentators, and we have aimed both to explain where some of the most difficult interpretive puzzles have arisen and how they are to be solved. We have shown how a truth theory can be employed in the pursuit of providing a meaning theory. We have shown how it solves a dilemma that Davidson develops in the critical phase of TM. We have shown that this project is continuous with that of TMLL. We have explained how passages in TM cited for the replacement theory fit with this project. And we have shown that the project we extract from TM is exactly the project that Davidson later pursues and attributes to his earlier self.

Notes

[1] A third less prominent suggestion is that he aims to reduce meaning to a special sort of truth conditions (Burge 1992: 20–1; Horwich 2005: 4, chapter 8). One version of the explication reading holds instead that Davidson aimed to replace, not reduce, meaning with a strong notion of truth condition.

[2] We have defended this position in Lepore and Ludwig (2003, 2005, 2007, 2011) and Ludwig (2002, 2011).

[3] Our discussion has focused on a central interpretive question about Davidson's project. Details about the use of an interpretive truth theory in giving a compositional meaning theory for natural languages can be found in Lepore and Ludwig (2005: part I, 2007).

References

Barwise, J. and 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. *Philosophical Review* 101(1): 3–51.

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

Cummins, R. (2002). Truth and meaning. In *Meaning and Truth: Investigations in Philosophical Semantics*, J.K. Campbell, M. O'Rourke, and D. Shier (eds). New York: Seven Bridges Press.

Davidson, D. (2001a). Introduction. In *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

———(2001b). Theories of meaning and learnable languages. In *Inquiries into Truth and Interpretation*, 2nd ed. Oxford: Clarendon Press. Original publication, 1965.

——— (2001c). Truth and meaning. In *Inquiries into Truth and Interpretation*, 2nd ed. Oxford: Clarendon Press. Original publication, 1967.

——— (2001d). Radical interpretation. In *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press. Original publication, 1973.

———(2001e). Reply to Foster. In *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press. Original publication, 1976.

——— (2001f). *Inquiries into Truth and Interpretation*, 2nd ed. Oxford: Clarendon Press. Original edition, 1984.

Ebbs, G. (2012). Davidson's explication of meaning. In *Donald Davidson on Truth, Meaning and the Mental*, G. Preyer (ed.). Oxford: Oxford University Press.

Glock, H.-J. (2003). *Quine and Davidson on Language, Thought, and Reality*. Cambridge: Cambridge University Press.

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

Katz, J.J. (1982). Common sense in semantics. *Notre Dame Journal of Formal Logic* 23:174–218.

Lepore, E. and Ludwig, K. (2003). Truth and meaning. In *Donald Davidson*, K. Ludwig (ed.). New York: Cambridge University Press.

——— (2005). *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Oxford University Press.

——— (2007). *Donald Davidson: Truth-Theoretic Semantics*. Oxford: Oxford University Press.

——— (2011). Truth and meaning redux. *Philosophical Studies* 154:251–277.

Ludwig, K. (2002). What is the role of a truth theory in a meaning theory? In *Meaning and Truth: Investigations in Philosophical Semantics*, D. Shier, J.K. Campbell, and M. O'Rourke (eds). New York: Seven Bridges Press.

——— (2011). Donald Davidson. In *Key Thinkers in the Philosophy of Language*, B. Lee (ed.). London: Continuum Press.

Quine, W.V. (1960). *Word and Object*. Cambridge: MIT Press.

Soames, S. (1992). Truth, meaning, and understanding. *Philosophical Studies* 65(1–2):17–35.

——— (2008). Truth and meaning: in perspective. *Truth and Its Deformities: Midwest Studies in Philosophy* 32:1–19.

Stich, S. (1976). Davidson's semantic program. *Canadian Journal of Philosophy* 6:201–227.

Tarski, A. (1983). The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, J. Corcoran (ed.). Indianapolis, IN: Hackett Publishing Company. Originally published, 1934.