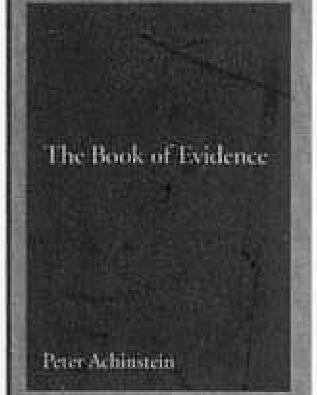

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO. 12, DECEMBER 2001

# TWO DOGMAS OF DAVIDSONIAN SEMANTICS*

In "Truth and Meaning," Donald Davidson[1] first formulated what was to become known as "Davidson's program." He proposed to elucidate the notion of natural-language meaning in general by showing how to construct a theory of meaning for a particular language, that is, a theory which would allow the interpretation of all the sentences of that language. Davidson's basic idea was to exploit a technique that Alfred Tarski invented in his endeavor to show how truth could be defined for a formal language. This led to the slogan that a theory of truth for a language could "serve as" a theory of meaning for that language. Davidson developed this idea in subsequent years, adding in particular his theory of radical interpretation, which was to explain how a theory of meaning for a particular language (in Davidson's sense) could be empirically confirmed.

It is my aim here to re-examine and criticize two doctrines which have become part of the Davidsonian program, but which are not essential to his original idea. They are the result, in my view, of a few

* I am grateful to Gabriel Segal for his extremely helpful comments on a recent version of this article. Parts of it have been presented on various occasions. One of these was the XVIth International Symposion of Philosophy, which was held in honor of Donald Davidson in October 1997 at the Instituto de Investigaciones Filosóficas of the UNAM in Mexico City, where I received helpful comments from Davidson, Mario Gómez-Torrente, John McDowell, Rupert Summerton, and other participants. Another occasion was Wolfram Hinzen and Hans Rott's workshop on Belief and Meaning at Regensburg in May 2000, where Johannes Brandl, Wlodek Rabinowicz, and others generously helped with comments. I would like to thank the audiences of these presentations for their comments and the organizers of these events for inviting me. Finally, I would like to thank David Papineau, Michael Potter, Mark Sainsbury, and Mark Textor for discussion and comment.

[1] *Synthese*, XVII (1967): 304-23; reprinted in Davidson, *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 17-36. Page references will be to the reprinted version (this holds for other cited articles reprinted there).

0022-362X/01/9812/613–35

wrong turns that the development of Davidson's program took during early debates in the 1970s. The first of these doctrines is the prima facie absurd view that a theory of meaning for a language does not say what any sentence of that language means. More precisely, this is the view that the target theorems of a theory of meaning for a language ought to take the extensional form of material biconditionals of the form '*s* is true if and only if *p*', so that the theorems of a theory of meaning do not *state* what the sentences of the language mean (or what their truth conditions or contents are), but rather "give the meaning" of sentences and allow us to interpret them if we have further information about these theorems. I shall call this the *biconditional doctrine.* The second doctrine is the view that the concept of truth plays a central explanatory role in Davidsonian theories of meaning for a language. I shall call this the *truth doctrine.*

I shall argue that the original reasons for adopting these two doctrines are flawed, and that there are, in fact, good reasons for not adopting them. Both doctrines are often uncritically accepted and have almost become part of the Davidsonian legacy. But, in fact, they are unjustified, and the main insights of Davidson's program do not depend on them. That is why, in my title, I have called them "two dogmas of Davidsonian semantics."

## I. DAVIDSON'S PROGRAM

In "Truth and Meaning," Davidson sets out to describe the form a theory of meaning for a particular language should take if it is to show "how the meanings of sentences depend upon the meanings of words" (*ibid.*, pp. 17, 23). This pretheoretical adequacy condition is dictated by the simple need to explain the fact that languages can be learned even though they contain an indefinite number of sentences, the meanings of which could not be learned one by one. Davidson proceeds by considering the form of the theorems that a theory of meaning would generate. He rejects theorems of the form '*s* means *m*', where the replacement for '*m*' is an expression referring to a meaning (he uses a form of the slingshot argument). He also rejects theorems of the form '*s* means that *p*', because...

> ...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 (*ibid.*, p. 22).

Davidson's point here is that, if we want a theory that entails theorems of the form '*s* means that *p*', then we need to know something about the logic of the expression 'means that': we need to know which inferences involving this expression are valid. But the expression is intensional, and the best available account of the logical properties of intensional contexts involves a notion of meaning or synonymy: within intensional contexts, the substitution of synonymous expressions preserves truth.[2] But if we are out to explain the notion of meaning in general, or in a particular language, we cannot employ a logic that presupposes the very same notion of meaning.

Because of these difficulties, Davidson then looks out for a different, extensional expression that is to fill the gap in theorems of the form '*s*...*p*', so that these theorems can be derived in a purely extensional axiomatic system. He says that:

> ...the success of our venture depends not on the filling but on what it fills. 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* (*op. cit.*, p. 23).

His suggested filling is 'is *T* if and only if', that is, some (initially uninterpreted) predicate 'is *T*' combined with the extensional sentential connective 'if and only if'. The adequacy of a theory of meaning for a language can now be captured by the requirement that the theory "entail all sentences got from" (*op. cit.*, p. 23) the schema '*s* is *T* if and only if *p*' when '*s*' is replaced by a structural description of a sentence of that language, and '*p*' by a translation of that sentence.

It then turns out that Tarski's recursive method of defining a truth predicate for a formal language provides a way of satisfying this adequacy condition on 'is *T*'. Thus Davidson's proposal is that at least one good way of explaining how the meanings of the sentences of a given language depend upon the meanings of words is to construct a Tarski-style recursive theory of truth which generates theorems of the form '*s* is *T* if and only if *p*' for every sentence of that language.[3] This is Davidson's basic idea: to exploit Tarski's recursive technique in order to do justice to the compositional requirements of a semantic theory. In subsequent works, Davidson developed the details of this approach by (a) showing how some recalcitrant natural-language constructions could be forced into the tight corset of a Tarskian recursive structure, and (b) explaining how a theory of meaning thus

[2] See Gottlob Frege, "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, c (1891): 25-50; English translation in M. Beaney, ed., *The Frege Reader* (Cambridge: Blackwell, 1997); see also Rudolf Carnap, *Meaning and Necessity* (Chicago: University Press, 1956).

[3] It is important to realize that Davidson never argued that this is the *only* way of constructing such a theory. All he says is that so far "we have no other idea how to turn the trick" (*op. cit.*, p. 23).

understood could be confirmed by empirical data (radical interpretation). Many philosophers and theoretical linguists have since joined Davidson in this effort.

## II. THE BICONDITIONAL DOCTRINE

The biconditional doctrine results from an obvious basic problem with Davidson's approach. Suppose we have constructed a Davidsonian theory of meaning for a language, that is, a theory that entails a theorem of the form '*s* is *T* if and only if *p*' for each sentence *s* of the language, such that what replaces '*p*' in each theorem is a translation of *s*. These theorems are, in common parlance, "interpretive" ("give/show the truth condition of a sentence"), but they do not *state* what the mentioned sentence means (or what its truth conditions are). They are just material biconditionals: all that is required for the truth of a material biconditional is that both sides flanking the biconditional have the same truth value. If I know that a given sentence is true if and only if snow is white, I do not thereby know that this sentence means that snow is white, nor do I thereby know that its "truth condition" is that snow is white. An easy way to see this is to consider the two sentences

(T1) 'Snow is white' is *T* if and only if snow is white.
(T2) 'Snow is white' is *T* if and only if grass is green.

If we wanted to say that by knowing (T1), I know that 'snow is white' means that (or has the truth condition that) snow is white, then we would have to say also that by knowing (T2), I know that 'snow is white' means that (or has the truth condition that) grass is green.

Davidson sets himself this problem in "Truth and Meaning," and his response is to say that in order for (T1) to give me information on what 'snow is white' means, I need to know in addition that (T1) has the status of a natural law, that is, is derivable from a truth theory that has been empirically confirmed in the right way (that is, through a process of radical interpretation) and is maximally simple.[4] (T2) is not so derivable—to derive it, one would need further, nonsemantic information on whether the proposition that 'snow is white' is true and the proposition that grass is green have the same truth value.

This, however, does not remove all problems. Consider:

(T3) 'Snow is white' is *T* if and only if snow is white and either grass is green or grass is not green.

[4] See *op. cit.*, pp. xiv, xviii, 26; and "Reply to Foster," in *Inquiries into Truth and Interpretation*, p. 174.

Just as (T2), (T3) is not interpretive. But unlike (T2), (T3) *is* derivable from a truth theory confirmed by a Davidsonian process of radical interpretation. At least it is thus derivable if (T1) is, because it is logically equivalent to (T1).[5]

Davidsonians normally respond to this new difficulty by introducing the notion of a "canonical T-theorem." A T-theorem is a metalanguage sentence of the form '*s* is *T* if and only if *p*' where '*s*' is a description of an object-language sentence and '*p*' is a metalanguage sentence that does not mention any object-language expressions. A canonical T-theorem, now, is a T-theorem that can be derived following a specified (canonical) procedure. Roughly, this procedure involves applying the semantic axioms concerning the syntactic constituents of *s* in an order that inverts the order in which *s* was constructed from its constituents, and then arriving at a T-theorem by repeated application of the rule of substitution of material equivalents.[6] All canonical theorems are interpretive and do not suffer from (T3)'s problem. Thus, if one knows that a theorem has been derived in the canonical way, one thereby knows that the sentence mentioned on one side of it is interpreted by the sentence used on the other side.

A modified response is given by Richard Larson and Gabriel Segal (*op. cit.*, §2.2.1-2; cf. Segal), who argue that a theory of meaning is supposed to model the knowledge which explains speakers' linguistic behavior, that is, it models the "semantic module." On their view, such a semantic theory consists of a set of semantic axioms and a set of rules of inference ("production rules")—rules that permit fewer inferences than classical logic. These inference rules are designed to permit only the derivation of interpretive T-theorems. This is an improvement on the standard response, because it avoids the detour of first formulating a theory with general logical inference rules, and then restricting the use of these rules by introducing the notion of a canonical proof. If a semantic theory is to model the information

[5] This difficulty was originally raised by John Foster in "Meaning and Truth Theory," in Gareth Evans and John McDowell, eds., *Truth and Meaning* (New York: Oxford, 1976), pp. 1-32. For excellent discussion of Davidson's way around the difficulty, see Gabriel Segal, "How a Truth Theory Can Do Duty as a Theory of Meaning," in U. Zeglen, ed., *Donald Davidson: Truth, Meaning, and Knowledge* (New York: Routledge, 1999), pp. 48-58. Segal thinks that (T3) could be ruled out on the grounds that it is not part of a maximally *simple* theory.

[6] Martin Davies spells out such a canonical proof procedure in his *Meaning, Quantification, Necessity* (New York: Routledge, 1981), p. 33. See also Christopher Peacocke's "Truth Definitions and Actual Languages," in Evans and McDowell, pp. 162-88; and Richard Larson and Segal, *Knowledge of Meaning* (Cambridge: MIT, 1995).

contained in the semantic module, such a gratuitous detour ought to be avoided (*op. cit.*, p. 559, footnote 14).

Both the standard and the modified response to the problem leave Davidsonians with the awkward biconditional doctrine: the theorems of a theory of meaning are material biconditionals which do not *state* what the sentences mentioned mean. They do, however, "give" the meaning, or truth condition, of the sentence, and additional information of what constitutes a canonical proof allows one to use such a theory for interpretive purposes. In Larson and Segal's formulation: humans have a semantic module which can be modeled as a T-theory, and humans treat the theorems that can be generated by the theory *as* interpretive (*op. cit.*, p. 39; Segal, p. 55).

It is surprising that the cumbersome biconditional doctrine has not received more critical attention. The view that a semantic theory, or a semantic module, does not, on its own, provide information on what sentences mean should have been highly suspect. Larson and Segal, who are the only Davidsonians who take the problem seriously, improve upon the standard version of the doctrine. But their claim that, by allowing only a restricted set of inference rules, one can avoid uninterpretive theorems should have led to further reflection: if the only T-theorems derivable in a semantic theory are interpretative ones, then it should have been possible to modify the theory in such a way that it generates genuinely meaning-specifying theorems of the form '*s* means that *p*' (or '*s*'s content is that *p*' or '*s*'s truth condition is that *p*').

It is easy to see that the unattractive biconditional doctrine is unnecessary, if one considers Davidson's original reason for introducing target theorems of the form '*s* is *T* if and only if *p*'. Davidson's reason, as mentioned above, was the intensionality of the expression 'means that'. He thought that the theorems of a theory of meaning could not take the form '*s* means that *p*' because in order to derive such theorems, he would need to know intensional logic, something he thought presupposed the notion of meaning. But both the standard version and the modified version of the biconditional doctrine provide an easy solution to this problem of deriving intensional theorems, or so I shall argue in the next section.

## III. HOW TO DERIVE INTENSIONAL THEOREMS

Let me explore in more detail how genuinely meaning-specifying theorems of the form '*s* means that *p*' could be derived in a semantic theory of Tarski-Davidsonian cut.

Consider the standard version of the biconditional doctrine first. It says that there is a canonical procedure following which one can derive, from the semantic axioms of the truth theory, all and only interpretive T-theorems. It would seem that this provides us with all we need to know about the intensional logic of 'means that'. Can we not simply add an inference rule that permits one to move from

(P) '*s* is *T* if and only if *p*' is a canonical T-theorem

to

(C) *s* means that *p*

Let me first make a general observation about the form of inference rules. Any inference rule for a theory of meaning $M$ would be formulated not in the language $L_M$ of $M$ itself but in a metatheoretic language. It would take the form of an inference schema with schematic letters ranging over expressions of $L_M$ (not the object language). The schema then indicates that certain inferential moves from sentences of $L_M$ to sentences of $L_M$ are permitted. This is not unusual: the inference rules of any theory can be formulated only metatheoretically.

The suggested inference schema is not a rule of this sort: it permits inferences from sentences of the form of (P) to sentences of the form of (C). The former, however, are not sentences of $L_M$, but metatheoretic. This is because instances of (P) mention $L_M$ sentences and ascribe to them the metatheoretically defined property of being a canonical T-theorem.

We need a proper metatheoretic rule that permits a move within $L_M$ from a T-theorem to its meaning-specifying counterpart if the T-theorem is canonically derived. The following formulation uses an adapted version of a standard notation for schematically stating inference rules:

(R)
$$\frac{\begin{array}{l}\vdots\ \text{(canonical derivation)}\\ s\text{ is }T\text{ if and only if }p\end{array}}{s\text{ means that }p}$$

Schema (R) shows that one may derive an $L_M$ sentence '*s* means that *p*'[7] if one has previously been able to derive an $L_M$ sentence '*s* is *T* if and only if *p*' in the canonical way. The three dots, together with

For convenience, I use ordinary quotes where corners would be appropriate.

the specification in brackets, give a metatheoretic instruction: they indicate that '*s* is *T* if and only if *p*' needs to have been canonically derived if the move to '*s* means that *p*' is to be legitimate.

This form of stating a rule of inference may appear unorthodox and therefore arouse suspicion. But a little reflection will show that it is not unorthodox. Consider, for example, a typical schematic formulation of the rule of conditional proof for the propositional calculus:

$$\text{(CP)}\quad \begin{array}{c} [p] \\ \cdot \\ \cdot \\ \cdot \\ q \\ \hline p \supset q \end{array}$$

Consider the role of the metatheoretic instruction here: the three dots between the bracketed '*p*' and '*q*' indicate that, if a formula *q* has been proved on the assumption that *p*, then one may infer '$p \supset q$'. Standard formulations of other rules of inference, such as reductio ad absurdum and constructive dilemma, involve even more complex metatheoretic instructions.[8] Thus, if these are legitimate formulations of inference rules (which I take it they are) then so is (R).

But if (R) is a legitimate form of characterizing a rule of inference, then all we need to make it work is a canonical procedure following which one can derive all and only interpretive T-theorems. As we saw, the standard move of Davidsonians in the face of Foster-type problems is to invoke the existence of precisely such a canonical procedure.[9] Thus, there is no reason why these Davidsonians could not introduce a rule like (R), thereby generating intensional, meaning-stating theorems and giving up the biconditional doctrine.

Now consider Larson and Segal's modified version of the biconditional doctrine. They claim that the inference rules "contained in the semantic module" allow the derivation of all *and only* interpretive T-theorems. If this is true, then there may be an even more direct way of deriving meaning-stating theorems: just add an inference rule that allows the move from any T-theorem to its meaning-specifying counterpart. Schematically:

[8] Compare, for example, Wilfred Hodges, "Elementary Predicate Logic," in D. Gabbay and F. Guenthner, eds., *Handbook of Philosophical Logic*, Volume I (Dordrecht: Reidel, 1983), pp. 1-131, especially pp. 29-30.

[9] See page 617 above and footnote 6.

$$\text{(R*)}\quad \frac{s \text{ is } T \text{ if and only if } p}{s \text{ means that } p}$$

Since a Larson-and-Segal-style semantic module employs inference rules that do not allow the derivation of uninterpretive T-theorems, we need no metatheoretic instruction that restricts this move to cases where the premise has been canonically derived (*op. cit.*, p. 40, footnote 15).

This is too simple, however. Not every properly derived $L_M$ sentence of the form '*s* is T if and only if *p*' is a T-theorem, that is, a theorem in which the right hand side '*p*' is replaced by an $L_M$ sentence which does not mention any object-language expressions. As a consequence, (R*) allows too much. For example, the theory will allow the derivation of sentences of the form "'*s* & *r*' is true if and only if *s* is true and *r* is true." But "'*s* & *r*' means that *s* is true and *r* is true" is false. Our inference rule should only be applicable to T-theorems. This needs to be included in the metatheoretic description of the premise in (R*). So the following might be a correct formulation of the rule for Larson and Segal's theory:

$$\text{(R**)}\quad \begin{array}{l} s \text{ is } T \text{ if and only if p} \\ \text{(where '}s\text{' is a description of an object-language sentence and '}p\text{'} \\ \text{is an } L_M \text{ sentence that does not mention object-language expres-} \\ \text{sions)} \\ \hline s \text{ means that } p \end{array}$$

Thus, both on the standard view and on Larson and Segal's, there is a way in which a theory of meaning can have intensional, meaning-specifying theorems.[10]

[10] Some theorists might require that all inference rules be encapsulated in the formal definition of a derivability notion. Adding a rule of inference, on this view, must take the form of adding a clause to the definition of 'derivable' or 'theorem'. This requirement is unproblematic, at least for my proposed modification of Larson and Segal's theory. Since it is already the case that only interpretive T-theorems are derivable in their theory, and since the definition of 'T-theorem' (as well as the corresponding metatheoretical instruction in (R**)) is in purely syntactic terms, (R**)'s work can be done by a clause like the following:

(R***) If '*s* is *T* if and only if *p*' is a T-theorem, then '*s* means that *p*' is also a theorem.

Things are less straightforward in the case of the proposal, made above, to add (R) to the T-theories of more standard Davidsonians (that is, those who operate with the notion of a canonical theorem). Unlike Larson and Segal, these theorists apply general deductive rules to their axioms, which then generate noninterpretive T-theorems. In order to do (R)'s work in a formal definition of derivability it may be necessary to start from scratch with a notion of derivability that does not allow

All this, however, should not be taken to suggest that we can do without T-theorems or that we can replace '— is true if and only if ...' early-on, or throughout, with '— means that ...'. In the derivation of a meaning specification of a sentence, application of (R) or (R**) is only the last step. The real work is done previously by a derivation of a T-theorem.

An example may illustrate why: consider the derivation of a T-theorem for some conjunctive sentence '$s$ & $r$'. One would first use the axiom for '&' to derive

(i) '$s$ & $r$' is $T$ if and only if $s$ is $T$ and $r$ is $T$

then one would use independently derived theorems of the form

(ii) $s$ is $T$ if and only if $p$
(iii) $r$ is $T$ if and only if $q$

to derive something of the form

(iv) '$s$ & $r$' is $T$ if and only if $p$ and $q$

using a rule of substitution of equivalents. No meaning-specifying theorems could play the role of (ii) and (iii) in this derivation. Of course, (R) could be applied to (ii) and (iii) directly, and would then yield the correct meaning specification for $s$ and $r$. But in order to derive the meaning specification for '$s$ & $r$', we need the original biconditional version of (ii) and (iii).

My proposal is therefore not intended as an objection to Davidson's view that "we have no other idea how to turn the trick" (*op. cit.*, p. 23) of formulating a theory that allows one to generate pairings of object-language sentences with their metalanguage translations from information about simple sentence constituents. Tarski's machinery has an indispensable role in the theory (until we find a different idea how to turn the trick). What I object to is the unreflected doctrine that a theory of meaning cannot say what sentences mean because we lack information about the logic of the intensional phrase 'means that'.

Thus, we know enough about the intensional logic of 'means that' in order safely to derive theorems of the form '$s$ means that $p$' as a final step. We *can* have a meaning theory that states what sentences mean. The cumbersome biconditional doctrine is an unmotivated

the derivation of noninterpretive T-theorems in the first place (that is, basically Larson and Segal's strategy).

dogma. We *can* after all utilize Tarski-style recursive machinery and still derive, in a second step, what sentences mean.[11]

## IV. DAVIDSON'S PROGRAM BEFORE THE TRUTH DOCTRINE

Once one has taken the step of recognizing that the biconditional doctrine is a dogma, it becomes easier to take an instrumentalist view of the role of the predicate 'is $T$' in some theorems of a theory of meaning: it enables the recursive machinery to generate interpretive T-theorems—and ultimately theorems of the form '$s$ means that $p$'. The important function of the predicate is that it allows us to generate theorems that pair object-language sentences with their metalanguage interpretations.

At the time of writing "Truth and Meaning," Davidson was promoting this view himself, as is shown by the remark I quoted above:

> ...the success of our venture depends not on the filling but on what it fills. 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 $s$ (*op. cit.*, p. 23).

Davidson continues in this vein later in the paper, when he discusses the problem of evaluative sentences. Evaluative sentences would constitute an obvious problem for Davidsonian meaning theories, if the predicate involved in the T-theorems were thought of as expressing the notion of truth. For it is controversial whether evaluative sentences (or the contents expressed by utterances of them) can be evaluated in terms of truth at all. But Davidson brushes any such worries aside, asserting, once again, that all that counts is whether the theory can generate the right theorems from its axioms, implying that it does not matter whether the predicate involved expresses the notion of truth:

> If we suppose questions of logical grammar settled, sentences like 'Bardot is good' raise no special problems for a truth definition. The deep differences between descriptive and evaluative (emotive, expressive, and so on) terms do not show here. Even if we hold there is some important sense in which moral and evaluative sentences do not have a truth value (for example because they cannot be verified), we ought not to boggle at "'Bardot is good' is true if and only if Bardot is good"; in a theory of truth, this consequence should follow with the rest, keeping track, as must be done, of the semantic location of such sentences in the language as a whole—of their relation to generalizations, their role in such

[11] This conclusion does not affect Davidson's reasons for rejecting theorems of the form '$s$ means $p$' where '$p$' *refers* to a meaning (that is, his slingshot argument).

> compound sentences as 'Bardot is good and Bardot is foolish', and so on. What is special to evaluative words is simply not touched: the mystery is transferred from the word 'good' in the object language to its translation in the metalanguage (*op. cit.*, p. 31).

These remarks show that Davidson did not object to the T-concept invoked in the T-theorems being distinct from the concept of truth—or at least from truth in any sense in which it cannot be applied to evaluative sentences. As long as a theory of meaning delivers the needed T-theorems which permit us to interpret speakers, it does not matter what exactly we mean by 'true' in the theory of meaning. Thus, at the time of writing "Truth and Meaning," Davidson did not subscribe to the truth doctrine—the view that the notion of truth plays a key explanatory role in theories of meaning for particular natural languages.

## V. THE EMERGENCE OF THE TRUTH DOCTRINE

Davidson changed his view on this when he developed the theory of radical interpretation.[12] It was considerations about the explanatory aims of Tarski's project as opposed to the explanatory aims of his own meaning theories which prompted this change of view and gave rise to the truth doctrine. Davidson expressed the new view in many of his works after 1973; for example, in the introduction to *Inquiries into Truth and Interpretation*:

> [W]hile Tarski intended to analyze the concept of truth by appealing (in convention T) to the concept of meaning (in the guise of sameness of meaning, or translation), I have the reverse in mind. I considered truth to be the central primitive concept, and hoped, by detailing truth's structure, to get at meaning.[13]

In "Radical Interpretation":[14]

> [A]ssuming translation, Tarski was able to define truth; the present idea is to take truth as basic and to extract an account of translation or interpretation (*ibid.*, p. 134).

And in "Belief and the Basis of Meaning":[15]

> Our outlook inverts Tarski's: we want to achieve an understanding of meaning or translation by assuming a prior grasp of the concept of truth (*ibid.*, p. 150).

And more recently in "The Structure and Content of Truth":

> The theory is correct because it yields correct T-sentences; its correctness is tested against our grasp of the concept of truth as applied to sentences....
>
> [I]t is our grasp of [the concept of truth] that permits us to make sense of the question whether a theory of truth for a language is correct (*op. cit.*, p. 300).[16]

Let me spell out some of the background of these remarks. In "The Concept of Truth in Formalized Languages," Tarski formulates an adequacy condition for definitions of truth for a language. This condition requires that an adequate definition of truth for a language *L* must entail for each sentence of *L* a theorem of the form '*s* is true in *L* if and only if *p*', where '*s*' is a "structural description" of that sentence and '*p*' is the translation of that sentence in the language in which the definition is stated. Thus, in order to apply this criterion for the adequacy of a definition of 'true in *L*', one needs to understand what it is for a sentence to be a translation of (or synonymous with) another sentence. Now, Davidson proposes to construct a structurally similar theory, also entailing theorems of the mentioned form, but with the aim of "giving" the meaning of each sentence of the language under discussion. In order to decide whether such a theory is adequate, Davidson cannot, obviously, presuppose knowledge of which sentences of the metalanguage are translations of which sentences of the object language, as this is what the theory is supposed to yield knowledge *of*. So Davidson needs to apply a different criterion of adequacy for his theories of truth (which are to serve as theories of

[12] See his remarks in the introduction to *Inquiries into Truth and Interpretation*, pp. xiv-xv, and "The Structure and Content of Truth," this JOURNAL, LXXXVII, 6 (June 1990): 279-328, especially p. 286, footnote 20.

[13] *Inquiries into Truth and Interpretation*, p. xiv.

[14] *Dialectica*, XXVII (1973): 313-28; reprinted in *Inquiries into Truth and Interpretation*, pp. 125-39.

[15] *Synthese*, XXVIII (1974): 309-23; reprinted in *Inquiries into Truth and Interpretation*, pp. 141-54.

[16] This quote may seem to permit the following reading: we need to grasp the concept of truth because we need to assess whether the theory, its theorems, are correct, that is, whether they are true. This, however, seems to me to be the wrong reading; for one of the most important lessons Tarski draws in "The Concept of Truth in Formalized Languages"—in *Logic, Semantics, Metamathematics* (New York: Oxford, 1956), pp. 152-278—is that we must distinguish 'true' as a predicate of the object language and 'true' as a predicate of the metalanguage. The concept of truth whose grasp is required according to the current reading would correspond to a predicate in the meta-metalanguage, as it gets applied to metalanguage sentences. Thus it would be distinct from the truth notion allegedly required to understand the T-theory itself and its theorems. Moreover, if this were the sense in which an understanding of truth was required, then this requirement would not be a special feature of meaning theories, but rather a requirement of *any* theory

meaning). How could it be decided whether a Davidsonian theory of truth is adequate?

Davidson does, in fact, have a very good answer to this question. It is given by his methodology of radical interpretation. But he also uses the following line of thought to convince himself that he needs to presuppose the notion of truth as a basic explanatory concept: assuming what it is for one sentence to translate another, "Tarski was able to define truth,"[17] so if one wants to explain meaning (translation) using the very same theory, one obviously needs to be in possession of the notion of truth already. In Paul Horwich's[18] words:

> ...we would be faced with something like a single equation and two unknowns...knowledge of the truth conditions of a sentence cannot simultaneously constitute *both* our knowledge of its meaning *and* our grasp of truth for the sentence (*ibid.*, p. 68).

Davidson concludes that the notion of truth must play a central explanatory role in the construction of theories of meaning.

The starting point of this line of thought is correct. But its conclusion is not. It is correct that Davidson needs a new criterion of adequacy, different from Tarski's. He cannot check whether all the theorems of a theory of meaning for a language are correct by checking whether their right-hand sides translate the sentence mentioned on their left-hand sides. But he does not need to check whether the theorems are correct by checking *immediately* whether their right-hand sides give sufficient and necessary conditions for the truth of the sentences mentioned on their left-hand sides. If he needed to be able to know necessary and sufficient conditions for the truth of all the sentences *immediately*, then the project of interpreting an unknown language would be hopeless.[19] Instead, what Davidson can do (and what he, in effect, proposes to do in his methodology of radical interpretation) is to check whether the theorems, if taken as interpretive, allow one to make good sense of the linguistic behavior of the speakers of the language.

[17] "Radical Interpretation," p. 134. This is not in fact an accurate description of what Tarski does in his article, as he did not define truth, but only defined truth in a particular formal language, as Davidson never gets tired of pointing out in later works—for example, in "The Structure and Content of Truth," and in "The Folly of Trying to Define Truth," this JOURNAL, XCIII, 6 (June 1996): 263-78.

[18] *Truth* (New York: Oxford, 1998, second edition).

[19] And in any case, if *that* were required, then it would not be *the* notion of truth that is presupposed, but the notion of truth in the language under discussion.

John McDowell,[20] in his reconstruction of the original Davidsonian project, exploits just this idea. A theory of meaning for a language ought ultimately to provide us with information that would be sufficient to interpret speakers of that language correctly. Now, a Davidsonian T-theory can take us part of the way. It can help us assign to every sentence a propositional content. But in addition to such a theory of content, we need a theory of illocutionary force. These two elements together form what McDowell calls a "bipartite" theory of meaning. If one knows such a theory for a language (and if one has sufficient time and patience) then it allows one to redescribe speakers' phonetic acts, that is, acts of emitting certain sequences of sounds, as propositional acts, for example, acts of asserting that, or asking whether such-and-such. According to McDowell, the acceptability of such a theory, as an empirical theory about the language's speakers, is measured by the extent to which these redescriptions allow us to make sense of the linguistic and nonlinguistic behavior of the speakers (*ibid.*, pp. 44-45). The working of such a theory, and the process of empirically confirming it, in no way depends on the interpretation of the predicate 'is *T*' that we are employing in our interpretive T-theorems and in generating them. Nothing prevents us from regarding 'is *T*' as a theoretical notion which is implicitly defined by the theory. Any (perhaps partial) coextensiveness between 'is *T*' and our ordinary notion of truth is something we discover afterward, it is not something we need to assume before we start the project. In McDowell's own words:

> The thesis should be not that [meaning] is what a theory of truth is a theory of, but rather that truth is what a theory of [meaning] is a theory of (*ibid.*, p. 47).

On McDowell's view, as long as we can make out a legitimate empirical methodology for constructing bipartite meaning theories, nothing obliges us to think of the T-notion employed in the theory of content as the notion of truth.

Again, rejecting the biconditional dogma puts one into a better position to see this. If we take a meaning theory for a language to yield predictions (theorems) of the form '*s* means that *p*', and not of the form '*s* is true if and only if *p*', then Davidson's motivation for the truth doctrine vanishes completely. Obviously, a radical interpreter is not required to know *immediately* whether the meaning specifying

[20] "Truth Conditions, Bivalence, and Verificationism," in Evans and McDowell, pp. 42-66, especially pp. 44-45.

theorems delivered by her theory of content are correct. She will test their correctness via further predictions these theorems allow us to make about speakers. Now, what further predictions these are will depend on how a theory of meaning *interacts* (to use a McDowellian expression) with other theories, that is, theories that predict which sentences speakers will utter given that they have a certain meaning, or in other words theories that explain why speakers utter sentences with a certain meaning.

Since the truth dogma is deeply entrenched, these abstract and general considerations are likely to meet with scepticism. But it should be agreed even by skeptics that the method by which a Davidsonian theory of meaning is to be tested empirically should be our ultimate touchstone for the truth doctrine: if the methodology of radical interpretation requires explanatory use of the notion of truth, then the truth doctrine is justified. If it does not, then the truth doctrine can be thrown onto the scrapheap of unjustified dogmas together with the biconditional dogma. In the next section, I shall therefore look in more detail at a methodology of radical interpretation in which the notion of truth makes no appearance whatsoever.[21]

## VI. RADICAL INTERPRETATION WITHOUT THE NOTION OF TRUTH

How can a Davidsonian theory of meaning for a particular natural language be empirically tested? In order to do this, we need to know more about the *observable* consequences of such a theory. What observable consequences could a theory have that states what the sentences of a language mean? Intuitively, if the theory is correct, speakers of the language will use certain sentences under certain conditions. For example, it would seem that, if some sentence meant that snow is white, then speakers would have a tendency to utter that sentence when they wish to get across that snow is white. If they do in fact have this tendency, then this confirms the theory; if they do not, it disconfirms the theory.

This commonsensical strategy is in principle correct. But we need to add a bit of theory before we can make some such strategy work. We need to clarify the connection between facts of meaning, as specified in the theorems of the modified Davidsonian theory, and the behavior of language users these facts should lead us to expect.

[21] Even though I follow McDowell in his assessment of the role of the concept of truth in these theories, I depart from him in the details of my account of radical interpretation, in particular in my treatment of illocutionary force (communicative function).

By assigning to each sentence a meaning, or truth condition, via a Davidsonian theory, one has not yet captured all meaning features relevant for communication. Sentences can also be classified as having various communicative functions. Some sentences serve to make assertions, others to ask questions or to issue commands. Combining a Davidsonian theory with a theory of communicative function, one can say that a Davidsonian theory specifies what the *content* of each sentence is, or what *proposition* it expresses, while the theory of communicative function specifies for the performance of which communicative act in relation to this content the sentence serves. For example, uttering the English sentence 'Sam smokes', one *asserts that* Sam smokes, while uttering the sentence 'Does Sam smoke?', one *asks whether* Sam smokes, performing a different communicative act on the same content.

Adding an assignment of communicative functions to a Davidsonian theory of content brings us closer to being able empirically to verify the now combined theory. The combined theory predicts that speakers perform certain actions of assertion, question, and so on when they utter sentences. But how do we test empirically whether a speaker is really performing such an action? We need a detailed account of the various linguistic acts, so that we can test the predictions the theory yields against general psychological assumptions concerning action.

The most promising accounts treat communicative acts as actions in intentional conformity with conventional rules and ultimately motivated by communicative aims. Philosophers like Ludwig Wittgenstein, J. L. Austin, H. P. Grice, John Searle, David Lewis, and Robert Stalnaker have pioneered this type of approach. Speakers aim to influence the beliefs of their audiences and audiences aim to acquire new information. Speakers know that audiences generally respond to utterances in conformity with the rules, and audiences know that speakers generally make utterances in conformity with the rules. This knowledge allows speakers and audiences to further their communicative aims by acting in accordance with the rules. There are characteristic rules for each kind of linguistic act (assertion, question, and so on).

Up to this point, most theorists of speech acts are in agreement. But they diverge considerably when it comes to the nature of the rules and the nature of the conventionality involved. I shall briefly sketch two approaches that seem to me to be particularly promising and comment on their suitability for a methodology of radical interpretation. For simplicity, I shall discuss only assertion.

The first approach is broadly along the lines of Lewis's[22] game-theoretical view of linguistic conventions and explains communicative acts directly in terms of belief-desire psychology. He defines a convention as a certain type of regularity in the behavior of a population of agents who face a recurring coordination problem. A coordination problem is a game-theoretical situation in which there are several equilibria on one of which the agents need to coincide. A convention is a solution to such a recurring problem: agents conform to a certain regularity of behavior because they expect the others to conform to the same regularity and it is in their interest to conform if the others conform. Thus conventional behavior can be explained directly in terms of a simple belief-desire psychology. In the case of linguistic conventions, in particular conventions regarding assertoric sentences, the relevant regularity might require agents (i) to utter a sentence assertoric of a content *p* only if *p*, and (ii) to respond to utterances of such a sentence by coming to believe that *p*.

This rule is only a first shot. It wrongly assumes that speakers who conform to linguistic convention are always truthful and sincere, and that audiences always believe what they are told. Clearly, speakers often inadvertently assert a content that *p* even though it is false that *p*. Moreover, speakers often deliberately assert *p* even though they do not believe *p* themselves. Similarly, audiences often fail to believe contents that have been asserted, or even fail to believe that the asserter believes what she has asserted. But let us assume that we can solve this problem by modifying Lewis's account in the following way.[23] The appropriate speaker regularity is that of asserting *p* only if one either (i) believes that *p*, or (ii) wants to give the impression that one believes that *p*, or (iii) wants to give the impression that one wants to give the impression that one believes that *p*, or.... And the corresponding audience regularity is that of responding to assertions of *p* by either (i) coming to believe that the utterer believes that *p*, or (ii) coming to believe that the utterer wants to give the impression that he believes that *p*, or (iii) coming to believe that the utterer wants to give the impression that he wants to give the impression that he believes that *p*.... As before, these regularities are conventions in Lewis's game-theoretical sense.

[22] *Convention* (Oxford: Blackwell, 1969), and "Languages and Language," in Keith Gunderson, ed., *Minnesota Studies in the Philosophy of Science, Volume VII* (Minneapolis: Minnesota UP, 1975), pp. 3-35, reprinted in Lewis, *Philosophical Papers, Volume I* (New York: Oxford, 1983), pp. 163-88.

[23] As I argue in my "Lewis, Language, Lust and Lies," *Inquiry*, xli (1998): 301-15.

We would then have a way of testing our combined theory against general psychological assumptions. Suppose the theory says that some sentence *s* is assertoric and expresses the proposition that *p*. If the theory is correct, then speakers will expect that audiences of *s* come to believe that the utterer either believes that *p* or wants to give the impression that he believes that *p* or.... If speakers' communicative aims together with these expectations provide a good explanation of utterances of *s*, then the theory is confirmed. The same goes, ceteris paribus, for audience behavior.

The advantage of such an account is that it offers an immediate integration of the combined theory into a general belief-desire psychology. Speakers have mutual knowledge of their conformity to certain regularities, and this, together with their communicative desires, motivates them, via ordinary instrumental reasoning, to engage in linguistic action.[24] On this model, a methodology of radical interpretation makes no explanatory use of the notion of truth. At least it does not so long as the notions of belief, desire, and their contents are independent of the notion of truth. But there is no prima facie reason to believe that these notions depend on the notion of truth—except perhaps Davidson's own reason for the truth doctrine, which I have discredited above.

The second approach I want to discuss might be called a "conversational approach to communicative action." It differs from the Lewisian approach in two ways. First, instead of Lewis's simple regularities concerning individual utterances of sentences, the conversational account states rules that specify the role of assertions in entire conversations, that is, in interconnected series of assertions by several agents. Second, the conversational approach makes these rules a matter of social norms, while Lewis's game-theoretical approach denies social norms or sanctions any role in linguistic convention. As an example for a conversational account, I shall here use the account of assertion proposed by Robert Brandom,[25] who argues that there is a

[24] A disadvantage of such an account is that the knowledge it attributes to speakers is at best implicit knowledge. Speakers do not go explicitly through the kind of instrumental reasoning this account suggests. The status of a combined meaning theory as a psychological hypothesis would therefore be unclear. Compare Stephen Laurence, "A Chomskian Alternative to Convention-Based Semantics," *Mind*, cv (1995): 269-301.

[25] "Asserting," *Noûs*, xvii (1983): 637-50, and *Making it Explicit* (Cambridge: Harvard, 1994). Another example would have been Stalnaker's pragmatic theory—for example, in "Assertion," reprinted in *Context and Content* (New York: Oxford, 1999), pp. 78-95, which has been further developed by Lewis, "Scorekeeping in a Language-Game," *Journal of Philosophical Logic*, viii (1979): 339-59. But, since Stalnaker always speaks of contents as truth conditions (which, in turn, he takes to be

system of social norms and rules that governs our linguistic interactions. Within this system, assertion has a central role: an assertion that $p$ counts (i) as an undertaking to justify $p$ if challenged to do so, and (ii) as issuing a license to use $p$ as a premise. Quite obviously, these rules rely on the normative vocabulary of social duties and licenses, which are ultimately explained in terms of notions such as authority and sanction. Thus a conversational account of assertion does not immediately yield reductive explanations of utterances in terms of the beliefs and desires of the utterer. All the same, the account provides an empirical test for bipartite theories of meaning for particular languages. If a general theory of action together with a candidate theory of meaning—now combined with a theory of communicative acts interpreted as social linguistic acts, can explain speakers' behavior, then this confirms the candidate theory.

Again, if a conversational account of communicative acts is used in our methodology of empirically testing a combined theory of meaning, there is no reason to believe that the notion of truth plays an explanatory role.[26] Thus, two promising theories of the communicative acts are at our disposal for use in a truthless methodology of radical interpretation. Nothing therefore prevents us from viewing the predicate 'is $T$' as contextually defined by the theory in which it occurs. There is no need to assume that the predicate expresses some pretheoretically familiar notion of truth, or that this pretheoretical grasp is required to endow the theory with explanatory power.[27] This confirms the conclusion of my earlier argument (in section v) that the truth doctrine is a dogma.

sets of possibilities), I prefer to use, for current purposes, an account that does not in any way appear to make explanatory use of the notion of truth.

[26] The Brandom account fares slightly better than the Lewisian as far as the needed independent account of the contents of belief and desire is concerned. There is no danger that the use of the notion of content of belief and desire reimports reference to the notion of truth, for Brandom already provides an independent account of this notion: the content of an assertion is constituted by its (material) inferential relations with others, that is, by what would count as a justification of the assertion and for what it would count as justification. To count as justification, of course, is another notion within the theory of social action, to be explicated, ultimately, in terms of authority and sanction.

[27] There are reasons to believe that 'is $T$' as contextually defined by a theory of meaning for a language $L$ will be coextensive with 'is true' in those areas where both can be applied. See McDowell, "Meaning, Communication and Knowledge," in Zak Van Straaten, ed., *Philosophical Subjects: Essays Presented to P.F. Strawson* (New York: Oxford, 1980), pp. 117-39, here p. 121; and Wiggins, "What Would Be a Substantial Theory of Truth?" in Van Straaten, ed., pp. 201, 203-04, and "Meaning, Truth-Conditions, Proposition: Frege's Doctrine of Sense Retrieved, Resumed and Redeployed in the Light of Certain Recent Criticism," *Dialectica*, XLVI (1992): 61-90. But this does not show that 'is $T$' and 'is true' express the same concepts. As McDowell emphasises, the coextensiveness of the two predicates, where the ranges over which they are defined overlap, should come as a *discovery*.

## VII. FINAL REMARKS

If I am right in claiming that the biconditional doctrine and the truth doctrine are dogmas, then the ultimate theorems of a theory of meaning do not need to take the extensional form '$s$ is $T$ if and only if $p$' and the use of the predicate 'is $T$' is merely an expedient in the recursive machinery of the theory. Pursuers of Davidson's program can coherently claim the notion of truth to have no explanatory significance in semantics. Where does that leave plausible-sounding slogans such as

(S1) The meaning of a sentence is its truth condition

and

(S2) To know the meaning of a sentence is to know under what conditions it would be true

Is the label "truth-conditional semantics" a misnomer?

The answer is that the terminology of truth conditions is indeed partly misleading. (S1) is informative only on the background of Davidson's biconditional dogma. It is correct insofar as it expresses the underlying Davidsonian insight that one can construct a theory of meaning for a language which explains the compositionality of meaning by exploiting recursive techniques first devised by Tarski to define truth in a formal language. But often (S1) is misleadingly used to suggest a deeper significance. (S2) is true only insofar as it expresses the insight that to know the meaning of a sentence is to know how to use it correctly. In some typical cases, it is correct to use a sentence only if its content is true, so knowing under what conditions it would be true (whatever that means) will often help one to use it correctly. But this insight is quite independent of Davidson's main idea in "Truth and Meaning," namely, to use a theory of a certain recursive structure as a theory of content for a language. It rather belongs to a theory of communicative (illocutionary) acts, which might state, for example, that one should assert that $p$ only if $p$ (= that one should assert that $p$ only if it is true that $p$).[28]

But is it not still too much of a coincidence that it should be possible correctly to use the predicate 'is true' in the intermediate

[28] See the discussion in the previous section for more accurate thoughts on assertion.

lemmas of the theory, if in fact any uninterpreted predicate would have served just as well?

The answer is that this is not a coincidence. It can be neatly explained by minimal (deflationist) assumptions about the function of the truth predicate without uncovering any deep link between truth and meaning.[29] Deflationists about truth claim that truth is not a mysterious property. All we need to know about truth is encapsulated in the way the truth predicate solves a simple syntactic problem. The problem arises, for example, when there are sentences (or propositions) that we can mention but not use (which we cannot make explicit). For example, many people are unable to state Fermat's theorem. Nevertheless, they can easily make reference to it by just using the expression 'Fermat's theorem'. (They often know that Fermat's theorem has only recently been proven.) Merely making reference to the theorem, however, does not yet allow one, for example, to assert the theorem or to use it as the antecedent of a conditional. This is where the truth predicate is useful. It is governed by the rule that by applying it to the name of any sentence (or proposition), one gets a new sentence which is equivalent to the sentence of which truth was predicated (or which expresses a proposition equivalent to the proposition of which truth was predicated). Some have expressed this by calling the truth predicate a prosentence forming operator.[30] For example, the sentence 'Fermat's theorem is true' expresses a proposition equivalent to Fermat's theorem. Another typical example is the sentence 'All of Davidson's doctrines are true', which is equivalent to the conjunction of Davidson's doctrines but much easier to state.

This simple function of the truth predicate explains why it should be possible to interpret the predicate used in T-theorems as the truth predicate. In generating our meaning-specifying theorems from semantic axioms, we need to pair structural descriptions of object-language sentences with their metalanguage translations. In temporarily "filling the gap," we ought to use the material biconditional, as that helps our derivations (via a rule of substitution of equivalents, see section III above). Since a structural description of a sentence cannot flank the biconditional, we need to complete it with some predicate 'is *T*' to form a sentence. Since the other side of the biconditional is that sentence's translation (if the theory is correct), it is no coincidence that interpreting 'is *T*' as the truth predicate makes the T-theorems come out true. For as we have seen, the truth predicate forms sentences that are equivalent to the sentence of which it is predicated. The usefulness of Tarski's methods in a compositional theory of meaning does not, therefore, indicate any deep connection between the notion of truth and that of meaning.

MAX KÖLBEL

New Hall/University of Cambridge

[29] As done by Michael Williams in his "Meaning and Deflationary Truth," this JOURNAL, XCVI, 11 (November 1999): 545-64, p. 557.

[30] For example, Brandom, in *Making It Explicit,* chapter 5.