Preprint: forthcoming in a special issue of *Argumenta* on the anniversary of Donald Davidson's birth in 2017.

Truth-theoretic Semantics and Its Limits

Kirk Ludwig
Philosophy Department
Indiana University

1. Introduction

Donald Davidson was one of the most influential philosophers of the last half of the 20th century, especially in the theory of meaning and in the philosophy of mind and action. In this paper, I concentrate on a field-shaping proposal of Davidson's in the theory of meaning, arguably his most influential, namely, that insight into meaning may be best pursued by a bit of indirection, by showing how appropriate knowledge of a finitely axiomatized truth theory for a language can put one in a position both to interpret the utterance of any sentence of the language and to see how its semantically primitive constituents together with their mode of combination determines its meaning (Davidson 1965, 1967, 1970, 1973a). This project has come to be known as truth-theoretic semantics.

My aim in this paper is to render the best account I can of the goals and methods of truth-theoretic semantics, to defend it against some objections, and to identify its limitations. Although I believe that the project I describe conforms to the main idea that Davidson had, my aim is not primarily Davidson exegesis. I want to get on the table an approach to compositional semantics for natural languages, inspired by Davidson, but extended and developed, which I think does about as much along those lines as any theory could. I believe it is Davidson's project, and I defend this in detail elsewhere (Ludwig 2015; Lepore and Ludwig 2005, 2007a, 2007b, 2011). But I want to develop and defend the project while also exploring its limitations, without getting entangled in exegetical questions.

We can distinguish two different projects in inquiries into meaning. The first is to give what I will call a theory *of* meaning, by which I mean a general account of the nature of linguistic meaning: what it is for words and sentences to be meaningful, what it is for particular words and expressions to mean what they do, and how this is bound up with our use of them. The second is to give what I will call a *meaning* theory, as opposed to a theory of meaning. A meaning theory, as I will be using the expression, is a theory for a particular language which admits of a division into a (finite number of) semantical primitives and (a possible infinity of) complex expressions, which can be stated in a finite form and which, in a sense to be made clear, enables us to specify for each of the sentences of the language what it means on the basis what its parts mean and how they are combined. A meaning theory, as I am thinking of it, is constitutively a compositional meaning theory. A further requirement on a meaning theory—the knowledge requirement—that knowledge of what the theory states put one in a position to understand any potential utterance of a sentence of the language on the basis of understanding the contained semantical primitives and how they are combined.

Truth-theoretic semantics is sometimes said to give us no more than a translation manual does (Soames 2008). The knowledge requirement shows that the sort of theory we seek cannot be given by a translation manual, for two reasons. First, one can understand a translation manual, that is, know what it states, without understanding either language. Second, a recursive specification of a translation from one language to another need not illuminate the compositional semantic structure of either. So if truth-theoretic semantics can fulfill the knowledge requirement, its content will go significantly beyond that of a translation manual. One of the questions I raise is whether truth-theoretic semantics can meet this requirement, and more broadly whether any meaning theory can.

The two projects, that of giving a theory of meaning and giving a meaning theory, are connected. There is now a long tradition in the philosophy of language of trying to shed light on the theory of meaning by way of reflection on how to construct and to confirm meaning theories for particular languages. This is analogous to trying to shed light on the concept of truth by way of constructing truth theories for particular languages. This is exemplified in Davidson's own project in the theory of meaning, the ur-project in this tradition, in which there are discernable two elements, what I have called elsewhere the initial and the extended projects (Lepore and Ludwig 2005). The initial project focuses on how to give a compositional meaning theory for a language by way of reflection on a truth theory for the language, taking for granted knowledge of what primitives mean. The extended project aims to shed light more generally on what it is for words to mean what they do by asking how one could confirm for a speaker a truth theory, on neutral evidence, that could be used for interpretation (Davidson 1973b, 1974, 1975, 1976).

My main focus in this paper is the initial project, that is, how to give a meaning theory for a natural language by way of constructing a truth theory for it, that is, truth-theoretic semantics. Toward the end of the paper I will come to some limitations of truth-theoretic semantics connected with the knowledge requirement, which point to the need to refocus on the theory of meaning more generally to achieve even as much insight into compositional meaning as it is the ambition of truth-theoretic semantics to provide.

The paper is organized as follows. Section 2 outlines the project of giving a meaning theory for a language and distinguishes two approaches. One takes the sentential complement 'that *p*' of the verb 'means' to refer to propositions and seeks to give a theory that directly issues in statements of what sentences mean of the form '*s* means that *p*'. The other is truth-theoretic semantics. I claim that associating propositions with every object language sentence is neither necessary nor sufficient for giving a meaning theory. That it is not necessary, for as much as can be done along the relevant lines, is to be shown in the main body of the paper. That it is not sufficient is shown toward the end of the paper in section 9, but as a kind of grace note to recognizing a limitation to the ambitions also of truth-theoretic semantics. Section 3 takes up the project of showing how appropriate knowledge about a truth theory enables one to interpret each object language sentence and understand its compositional semantic structure. The account in this section follows the main lines, with some refinements, of the account in (Lepore and Ludwig 2007a), though I postpone two issues about its adequacy until sections 4 and 5, where I confront an important recent objection, and offer a response that calls for a further revision. Section 4 takes up two questions about the approach. The first is whether the sort of knowledge said in section 3 to suffice to enable one to use a truth theory to interpret its object language is sufficient (Hoeltje 2013). The second is whether ultimately (for other, more subtle reasons) the approach illuminates meaning only by relying on antecedent competence in expressions known to be systematically related in meaning to expressions in the object language. The bulk of section 4 takes up the first question. Section 5 takes up the second question and argues that there is an interesting feature of how 'that'-clauses function in explicit statements of meaning that shows that antecedent competence in a language plays an ineliminable role in how they give us insight into the meanings of the sentences they are about. Section 6 returns to the question whether assigning propositions to sentences may avoid the difficulties surveyed and argues that it rather reinforces the lesson. My conclusion is that there is a certain sense in which no propositional knowledge of a truth theory or even a more direct meaning theory suffices for understanding the object language. Section 7 is a short conclusion that suggests that to break out of the circle of language that traditional approaches leave us in we need to relate words and sentences to the roles they are supposed to play in our communicative activities described in more fundamental terms.

2. Meaning theories

The most straightforward way to provide a meaning theory would seem to be to provide an axiom for each primitive expression in a language *L*, which (in some straightforward sense) gives its meaning, from the totality of which one could derive formally a specification of the meaning of any sentence of the language relative to a context of utterance—a specification, moreover, grasp of which enables us to understand the sentence, while the mode of derivation enables us to understand how our understanding of the sentence rests on our understanding of its significant parts and their mode of combination. A target theorem (for a declarative sentence) could take the following form (M).

(M) For any speaker *s*, and time *t*, ϕ means in *L*, taken relative to *s* at *t*, that *p*.

We allow '*p*' to be replaced by an open sentence containing as free variables '*s*' and '*t*' which will then be bound by the initial quantifiers in (M). Call theorems of this form M-theorems. So far, so good. Now we have a divergence between at least two different approaches.

First, a natural thought is to take 'that *p*', when '*p*' is replaced by a closed sentence, to be a referring term, and see the goal as being to formulate a theory that enables us to associate with each sentence of *L* and each context of utterance a referring term of the form 'that *p*' so as to enable us to understand the sentence as uttered in that context. Let us give these referents a uniform name: propositions. For each sentence, we want to be able to derive M-theorems from axioms attaching to its primitive components. This amounts to associating a referring term with each sentence derived from axioms governing its parts. Thus, it is natural to take the axioms governing its primitive components to be relating its parts to objects in such a way that we are in a position to understand those parts, and the proposition to be a structured complex of those parts, the structure paralleling in some way the mode of combination of the primitive expression in the object language sentence. In this way we make provision for the use of the power of quantificational logic in deriving theorems from axioms and to show how the meanings of sentences depend on the meanings of their parts.

Putting aside how the technical details are to be worked out, the question arises what role the assignment of objects to expressions and sentences in the language is doing in meeting the primary goal of the meaning theory, as specified above, and in particular the knowledge requirement. The answer is that associating entities with every expression in the language is, first, not necessary in order to provide a meaning theory for a language, to the extent to which this can be done, and, second, not sufficient either.

I will take up first the claim that it is not necessary, by introducing the second approach, truth-theoretic semantics: the project of formulating a meaning theory for a language *L* (focusing for the moment on its declarative sentences) in terms of a certain body of knowledge we can have about a truth theory for the language. We will return to the second claim in a roundabout way through considering ultimately a striking limitation on the ambitions of truth-theoretic semantics.[1]

3. Truth-theoretic semantics

Truth-theoretic semantics aims to exploit the recursive structure of a Tarski-style truth theory in pursuit of giving a compositional meaning theory. However, a Tarski-style axiomatic truth theory is not a meaning theory. It blandly states conditions materially necessary and sufficient for the truth sentences of an object language. Davidson's idea was that out of an extensional truth theory for a language, we can squeeze the elixir of meaning, if it has certain properties, and we know that it does. I develop that idea in this section, in a way similar to, if not quite the same, way that Davidson did. For illustration, I introduce a simple axiomatic truth theory for a language *L*, [TRU], without quantifiers or context sensitivity, whose sentences are all declarative (for extensions to quantifiers and to non-declaratives see (Lepore and Ludwig 2007a, chs. 3 and 12)). All of the central issues that concern us can be raised in connection with even this very simple theory. In the following, for convenience I use '$x + y + \dots$ ' to mean 'the concatenation of x with y with … in that order'. 'N' ranges over names and 'S', 'S_1' and 'S_2' over sentences of the object language.

[1] Recently Greg Ray (2014) has offered a third way that takes neither Davidson's indirect route through a truth theory nor quantifies over propositions, but rather, by ascending to a meta-metalanguage, seeks to give a recursive theory that generates theorems about the truth of M-theorems (MnT-sentences, of the form '"*s* means that *p*" is true'), and then by semantic descent to arrive at M-theorems. So if one could only generate recursively the set of MnT-sentences for the language, the idea is that one would be able to grasp an explicit statement about the meaning of every object language sentence. Ray shows cleverly how to formulate a recursive theory to generate the appropriate class of MnT-sentences. But the approach, as sketched in the paper, founders on the bit of reasoning that gets us from the MnT-sentences to the M-sentences. The reasoning is illustrated in the following:

(1) '"L neige est blanche" means that snow is white' is true
(2) If '"L neige est blanche" means that snow is white' is true, then 'La neige est blanche' means that snow is white.
(3) Hence, 'La neige est blanche' means that snow is white.

The crucial bit in getting from what (1) states, which is output of the theory, to what (3) states, which is what we want to know, goes by way of (2), which seems simply to be an instance of semantic descent. But this presupposes that the meta-metalanguage and the metalanguage are the same, and so effectively presupposes that the theorist already understands the metalanguage. However, in general knowledge of the truth of a sentence does not give us knowledge of the language in which it is stated, and the theory that Ray presents, in the meta-metalanguage, is about sentences in the metalanguage, and could be stated in a language other than the metalanguage. Thus, knowledge of the theory (of what it states) is not sufficient for knowledge of the metalanguage in which meaning statements are expressed, which means that it does not suffice for knowledge of what (2) states. This point has been ably made now in print by (Hoeltje 2016).

Truth theory [TRU]

1. 'Claudine' refers to Claudine
2. 'Robert' refers to Robert
3. For any name *N*, *N* + 'dort' is true in *L* iff what *N* refers to is sleeping.
4. For any names N_1, N_2, N_1 + 'aime' + N_2 is true in *L* iff what N_1 refers in *L* to loves what N_2 refers to in *L*.
5. For any sentence *S*, 'Ce n'est pas le cas que' + *S* is true in *L* iff it is not the case that *S* is true in *L*.
6. For any sentences S_1, S_2, S_1 + 'et'+ S_2 is true in *L* iff S_1 is true in *L* and S_2 is true in *L*.

Rules of Inference

Universal Instantiation (UI): from a universally quantified sentence any instance may be inferred.
Substitution (S): from any sentences of the form '*t* refers to *y*' and any sentence of the form S(what *t* refers to in *L*), S(*y*) may be inferred.
Replacement (R): S(*y*) may be inferred from '*x* iff *y*' and S(*x*).

Our criterion of adequacy for the truth theory is Tarski's Convention T, which requires that

The truth theory entails all instances of the schema (T)

(T) φ is true in *L* iff *p*

in which φ is replaced by a structural description of an object language sentence as composed out of its significant parts and '*p*' is replaced by a metalanguage sentence that translates it.

A canonical truth theorem of [TRU] is any theorem derived from the axioms using only (UI), (S), and (R) whose last line is of the form (T) and in which no semantic vocabulary of the metalanguage remains (i.e., 'is true in *L*'). We call such a proof a *canonical proof*. It is clear that a canonical proof draws only on the content of the axioms in proving a canonical theorem.

What knowledge about [TRU] would enable us to know that it meets Convention T? We stipulate that

(i) in each reference axiom, the name used on the right of 'refers to' translates the name mentioned on the left
(ii) in each predicate axiom, the predicate used in the meta-language in giving truth conditions for the object language sentence translates the object language predicate.
(iii) in each recursive axiom, the logical connective used in the meta-language to give the truth conditions for the object language sentence translates the logical connective in the object language

We will say that if the theory meets this condition, it meets Convention A. We will say that a truth theory that meets Convention A is interpretive. It is clear that given the rules of inference and that [TRU] meets Convention A, for each sentence of the object language its canonical theorem is such that the metalanguage sentence on the right hand side translates the object language sentence for which it is used to give truth conditions, and, hence, that [TRU] satisfies Convention T.

The step to seeing how to transform these materials into a meaning theory is accomplished by restating Convention T.

> The truth theory entails all instances of the schema (T)
>
> (T) φ is true in *L* iff *p*
> (M) φ means in *L* that *p*
>
> such that the corresponding instance of schema (M) is true.

At this point, given that [TRU] satisfies Convention A, and, hence, Convention T, we can add another valid rule of inference (cf. (Davidson 1970; 2001, p. 60)):

> [Transference] 'φ means in *L* that *p*' may be inferred from a canonical truth theorem of the form 'φ is true in *L* iff *p*'.

We will call any theorem so derived a canonical meaning theorem.

What this shows is that we can use [TRU] to arrive at a specification of the meaning of each sentence of the object language. This does not, however, make [TRU] a meaning theory for the object language. It is still just a truth theory. It doesn't state anything itself about what sentences in the object language mean, as opposed to the conditions under which they are true. But if it is the right sort of theory, and we know that it is, and understand it, then, it seems, we can *use it* to specify what each sentence means, in a way that makes it scrutable to us, and which, by the proof of the relevant theorem, shows how the semantical primitives in it contribute to fixing its truth condition determining meaning.

We characterized a meaning theory as a body of knowledge that puts one in a position to understand any sentence of a language on the basis of understanding the contained semantical primitives and their manner of combination in the sentence. With this in mind, we will identify the meaning theory for the object language with what body of knowledge we need to have about [TRU] that puts us in a position to use it for this purpose.

In particular, if we know [K-TRU],

> [K-TRU]
>
> (i) What the axioms of [TRU] are, as stated in (1)-(6).
> (ii) What each axiom states and of each that it states what it does.
> (iii) That [TRU] is interpretive, i.e., meets Convention A.
> (ii) The rules of inference UI, S, R, and T.

(v) What a canonical truth theorem is.

then (it seems) we are in a position to infer for each sentence of the object language a meta-language sentence that explicitly states what the object language sentence means, on the basis of a proof that traces out, at each step, the contribution of each object language expression to fixing the truth conditions of the sentence to which it contributes, on the basis of reference or truth conditions given using a term synonymous with it. We can thus see what the contribution is of each semantical primitive to the interpretive truth conditions of the sentences in which it occurs on the basis of what it means. Thus, we can treat the body of knowledge characterized by [K-TRU] as a meaning theory for *L*. Here again the propositional knowledge stated in (ii) is to play the crucial role of giving us knowledge of the meaning of the axioms of the truth theory and so of the language of the metalanguage (i.e., of the truth theory). If this is correct, it shows that our meaning theory goes beyond a translation theory because one can grasp what the translation theory states without understanding either language or the compositional semantic structure of their sentences.

That completes the basic account of the goals and methods of truth-theoretic semantics. I close this section of the paper with four remarks.

(1) First, it is clear that on this way of understanding the project of truth-theoretic semantics, the truth theory is not identified with a meaning theory. The meaning theory is rather a body of knowledge about the truth theory. It is not an objection to the project then that the truth theory itself cannot do the duty of a meaning theory.

(2) Second, it is clear that the project is not to replace the investigation of meaning with the investigation of truth on the grounds that the concept of meaning is too confused for use in a properly scientific investigation of language, but rather the pursuit of a traditional project by means of a certain kind of indirection.

(3) Third, it is clear that it is not the goal of truth-theoretic semantics, as here conceived, to effect a kind of reduction of meaning to truth conditions, in any sense.

(4) Fourth, the approach shows how to achieve the aims of a meaning theory with no more ontological resources than are required for a reference theory, and, hence, that the introduction of entities to assign to every significant expression of the language is not necessary in order to give a meaning theory for a language.

These are, I believe, all observations which conform to Davidson’s own understanding of his project (Ludwig 2015). The distinction between the truth theory and what we can know about it that enables us to use it to interpret a language is drawn clearly in “Reply to Foster” (Davidson 1976). However, Davidson did not think that a statement of what we can know would count as a theory, as he says on the last page of “Reply to Foster,” largely because of his commitment to analyzing ‘states that’ (or alternatively ‘means that’) paratactically (Davidson 1968), which is required in spelling out [K-TRU] (ii). When we put aside the commitment to the paratactic analysis, though, there would seem to be no barrier to stating what propositional knowledge would suffice explicitly. (We will return below in sections 7-9 to the question whether we have achieved everything we want.) Another departure from Davidson's own development is the introduction of Convention D as a constraint on an interpretive theory. As noted in the introduction, Davidson's project was not limited to formulating a compositional meaning theory for a language. When the project is to understand how we understand complex expressions on the basis of their components and combination, we can help ourselves to knowledge of what the primitive expressions mean. Davidson sought illumination also of what it was for primitive expressions to mean what they do. The method was to describe empirical constraints on truth theory sufficient to ensure that the theory could be used for interpretation in the form of the requirement that it be confirmable for a speaker from the standpoint of a radical interpreter—an interpreter who starts ultimately with only behavioral evidence about a speaker interacting with his environment and others. If the constraint were adequate to guarantee not merely the right outputs, but also that any theory so confirmed would provide insight into how the sentences of the language were understood on the basis of understanding of their contained primitives and mode of combination, then it would suffice for the theory to meet Convention D as well. Since my interest is in the truth theory as a vehicle for a meaning theory, I stipulate that part of what we need to know about the truth theory is that it meets Convention D and what each axiom means.

5. Is the relevant body of knowledge really adequate?

What I have hoped to do up to this point is

> (1) to explain how truth-theoretic semantics is to be understood as a pursuit of a perfectly traditional project, that of constructing meaning theories for particular languages;
> (2) to show that it is mistake to suppose its work can be done by a translation theory.

Now I want to return take up two important questions about it.

> (i) Is the body of knowledge I have claimed to be sufficient really so?
>
> (ii) Even if it is sufficient in some straightforward sense, and granting that it is not equivalent in any straightforward sense to a recursive translation theory, might there not yet be a sense in which the illumination of what object language expressions and sentences mean rests in part essentially not upon propositional knowledge but upon non-propositional understanding of the metalanguage, that is, antecedent competence in expressions known to be systematically related in meaning to expressions in the object language?

I address the first question in this section and the second in the next. The first question is prompted by an objection due to Miguel Hoeltje (2013), namely, that what is stated in [K-TRU] (repeated here) does not suffice to understand the truth theory.

[K-TRU]

(i) What the axioms of [TRU] are, as stated in (1)-(6).
(ii) What each axiom states and of each that it states what it does.
(iii) That [TRU] is interpretive, i.e., meets Convention A.
(iv) The rules of inference UI, S, R, and T.
(v) What a canonical truth theorem is.

The meaning theory involves three levels of languages. There is the object language. There is the language of the truth theory, the metalanguage, and then there is the language in which the meaning theory is stated, the meta-metalanguage. The goal was in part to state something in the meta-metalanguage sufficient to understand the metalanguage. This is important because (a) both the canonical truth theorems and canonical meaning theorems are in the language of the truth theory and (b) the canonical proofs, which are to illuminate the compositional structure of object language sentences, are stated in the meta-metalanguage. The objection that Hoeltje makes is that (ii) in [K-TRU], which is intended to turn the trick, is not sufficient. Let's take an example of the sort of knowledge that gives us. I'll vary the language of the truth theory from English to Serbian (in the Cyrillic alphabet) to bring out the point.

[*] За свако име н, н⁀“dort” је истинито у л акко оно на шта н реферира спава'
means that for any name *N*, *N* + 'dort' is true in L iff what *N* refers to is sleeping.

The worry is this: how are we supposed to know anything about the relevant syntactic/semantic structure of axiom, and hence of the metalanguage sentence, from basically getting a statement of its meaning as a whole? And if we don't know that, how can this give us knowledge of the metalanguage, i.e., the language of the truth theory, which we grant is essential to using it in the way intended?

As a first remark about whether what is stated in [K-TRU] is sufficient, it is worth noting that if we have an explicit statement of the form

A means that *p*

for each of the axioms of the truth theory, then *we can state a truth theory in the meta-metalanguage for the object language*. It just consists in the list of statements that replace '*p*' in this form. We would of course also have to introduce inference rules corresponding to those for the metalanguage and define a canonical truth theorem and canonical meaning theorem for the meta-metalanguage. This is straightforward, but isn't yet included in what we would know in knowing the axioms or in knowing what else is stated in [K-TRU]. This looks like it would suffice. So just the knowledge stated in (ii) goes a long way toward what is needed, so far that just a bit more seems to gets us the rest of the way.

However, this is not how I envisioned it going. So let me return to the original idea, which was not adequately spelled out in (Lepore and Ludwig 2007a). First of all, we run proofs on the truth theory as a syntactic object. So we should state the rules in the meta-metalanguage. This was not made explicit in [K-TRU]. Furthermore, the rules are stated in terms of syntactic/semantic categories that apply to metalanguage terms (names, n-ary predicates, connectives, quantificational determiners, and so on), so it presupposes a recursive syntax for the metalanguage which sorts terms into the categories necessary for the description of the structure of sentences in metalanguage in terms of the types of semantically primitive expressions in the language and how sentences are systematically built up out of them. Armed with this information, we will be able to parse the structure of the axioms of the truth theory, and given the knowledge stated in (ii), see how to interpret connectives, determiners, predicates and so on. In fact, we also omitted to include in the statement of the truth theory a statement of the corresponding information about the object language. That should be included in the statement of the truth theory, and bundled into (ii). So what we need to make explicit that we had not is that we also know a lot about the syntactic structure of the metalanguage, and that the truth theory should itself include a recursive syntax for the object language.

For example, if we know that in

За свако име н, н⁀“dort” је истинито у л акко оно на шта н реферира спава

‘н’ is a variable, that ‘акко’ is a logical connective, that ‘За свако име’ is a restricted quantifier, that ‘За свако’ is a quantificational determiner, that ‘име’ is a common noun, that enclosing an expression in the left and right double quotation marks forms a name, that for metalinguistic variables *v* and *u*, *v* + ‘⁀’ + *u* is a restricted quantifier, and how the sentence is constructed out of its parts, which will determine scope relations, etc., we’re in a position to interpret basically each word in it given what it means as a whole, on the basis of knowledge of its syntactic structure and the knowledge of the syntactic structure of the sentence that gives its meaning in [*] and the meaning of that sentence, which is a sentence of our meta-metalanguage. Thus, given that we know that ‘За свако’ is a quantificational determiner, that ‘акко’ is the main connective, and that ‘н’ is a variable, we can infer that ‘За свако’ means *for any*, that ‘име’ means *name*, that ‘н’ is a variable that takes names as values, that ‘н⁀“dort”’ means *the concatenation of N with “dort”*’, that ‘је истинито у л’ means *is true in L*, that ‘акко’ means *iff*, that ‘на шта н реферира’ means *what N refers to* (which will be invariant across many axioms, as will ‘је истинито у л акко’, and ‘За свако име н’, which also helps us to identify the function of these phrases in the metalanguage), and finally that ‘спава ’ means *is sleeping*. I conclude that this challenge can be met, and [K-TRU] is to be amended in the ways indicated.

8. The limits of truth-theoretic semantics

Let’s turn to the second question:

> (ii) Even if knowledge of the sort indicated is sufficient in some straightforward sense, and granting that it is not equivalent in any straightforward sense to a recursive translation theory, might there not yet be a sense in which the illumination of what object language expressions and sentences mean rests in part essentially not upon propositional knowledge but upon non-propositional understanding of the metalanguage, that is, antecedent competence in expressions known to be systematically related in meaning to expressions in the object language?

I think the answer to this question is ‘yes’, and that it helps to bring out something important about the limits of truth-theoretic semantics, as it has been explained here.

Let’s grant knowledge of the language of the truth theory and think about how what we know about it helps us to use it to interpret object language expressions and to gain insight into the logico-semantic structure of object language sentences. First of all, knowing that the theory satisfies Convention A enables us to read off from axioms what object language expressions contribute in context to the meaning of a sentence. But the knowledge it gives us of both the content of the object language expression and its semantic role is clearly relative to our prior grasp on the corresponding content of the metalanguage expression and its semantic role. Let’s take five sorts of axioms as examples I shift to a context sensitive reference axiom and satisfaction predicate for generality.

A1. For any speaker *s*, time *t*, for any *x*, ‘je’ refers(*s*, *t*) to *x* iff *x* = *s*.
A2. For any function *f*, speaker *s*, time *t*, referring term *N*, *f* satisfies(*s*, *t*) *N* + ‘dort’ in *L* iff what *N* refers to is sleeping at *t*.
A3. For any function *f*, speaker *s*, time *t*, variable *v*, *f* satisfies(*s*, *t*) *v* + ‘dort’ in *L* iff f(*v*) is sleeping at *t*.
A4. For any function *f*, sentences S_1, S_2, *f* satisfies(*s*, *t*) S_1 + ‘et’+ S_2 iff *f* satisfies(*s*, *t*) S_1 in *L* and *f* satisfies(*s*, *t*) S_2 in *L*.
A5. For any function *f*, predicate *P*, variable *v*, *f* satisfies(*s*, *t*) ‘(Chaque’ + *v* + ‘)’ + *P* iff every *v*-variant *f′* of *f* is such that *f′* satisfies *P*.

A1 is the most informative of these. Antecedent understanding of ‘refers’ tells us what the category is, and (given Conv A) we know that the clause expresses a rule of meaning for determining the referent of the expression without using an expression the same in meaning with it. But in the case of A2-A5, it is clear that understanding the content and semantic role of the object language expressions depends on our antecedent understanding of the metalanguage sentence used to give satisfaction conditions together with knowledge that the theory meets Convention A. Similarly, our understanding of the contribution of each expression to fixing interpretive truth conditions of a sentence on the basis of its meaning, as exhibited in how its axiom enters into a canonical proof of a canonical theorem for it, likewise relies on our understanding of the metalanguage expression used to give satisfaction conditions for it and our knowledge that the theory satisfies Convention A.

There is a sense, then, in which the theory shows something about the object language, in light of our understanding of the metalanguage and knowledge that the theory meets Convention A, that the theory does not say. This, I think, turns out to be inescapable in giving a meaning theory of the kind we have set as our goal here. I’ll say more in support of this in a moment. This is not to say that the theory together with knowledge that it meets Conv A is not informative, of course, but only to say that the mode by which we gain information about the object language rests on prior understanding of metalanguage terms and what information Conv A gives us about their relation to object language terms for which reference and satisfaction conditions are being given. This is what survives of the objection that truth theoretic semantics could be replaced by a translation theory, namely, that it presupposes prior knowledge of a language whose terms are understood to be the same in meaning as object language terms.

But if all of this is right, if the illumination the theory provides does rest essentially in part on antecedent understanding of a language, then in what sense have we stated something in [K-TRU] that puts us in a position to understand any potential utterance of an object language sentence and to understand its compositional semantic structure? In a perfectly ordinary sense of 'body of knowledge', [K-TRU] (as modified along the line indicated at the end of the last section) does state a body of knowledge that suffices, for we can state that one needs to know that *p*, that *q*, etc., and we agree that if all of that is true of someone then he is in a position to interpret object language sentences. But this, I think, turns out to only superficially satisfy the requirement we had in mind. For what we had in mind was that we could state a body of propositional knowledge that suffices which is essentially independent of knowledge of any language. And, on closer examination, this condition has not been satisfied.

I want to take this up in connection with the final output of the theory, which consists in explicit statements about what object language expressions mean relative to contextual parameters. For I think this is the crux of the issue, how M-sentences sentences convey to us what a sentence means, relative to a context. Consider (M-S) for example.

> (M-S) For any speaker *s*, time *t*, 'Je dort' means(*s*, *t*) in *L* that *s* is sleeping at *t*.

Surely here we have got something that simply states what the sentence means! This at least does not rely on antecedent understanding of the metalanguage sentence and knowledge that it interprets the object language sentence. But *in fact it does*. We have not escaped reliance on antecedent understanding of the metalanguage.

It is easier to see this by instantiating (M) to a particular speaker and time, KL at T.

[KL] 'Je dort' means(KL, T) in *L* that KL is sleeping at *t*.

You understand the sentence used in the complement. And given that you understand it, and that you know that

> '*x* means(*s*, *t*) in *L* that *p*' is true in English iff *x* taken relative to *s* at *t* in *L* translates '*p*' in English,

you are in a position to understand the object language sentence taken relative to KL and T. But suppose that you did not understand the complement sentence. Then it is clear that you would not understand 'Je dort' taken relative to KL and T. So in fact the explicit statement of meaning relies for its effect on your understanding a sentence and understanding it to give the meaning of another sentence in a context.

To this it might be said:

> Yes, but the trouble is just that if you didn't understand the complement sentence in KL, you would not grasp what proposition was expressed by the whole sentence, and so of course we would not expect you to be in a position to understand the object language sentence. This does not show at all that what is *stated* is not

> sufficient. What is stated is that the meaning is a certain proposition. And if you grasp that you grasp the meaning of the target sentence. It does not show that the illumination comes only in the manner you sketch, i.e., via understanding of the complement sentence and knowledge that it translates the object language sentence—again, you just have to know what proposition it refers to! And while it is still true that we have to understand the complement to understand what the sentence states, that is just a reflection of the fact that to understand what someone is stating we have to understand the sentence he uses to state it.

But, as I will now argue, this response is inadequate and appeal to propositions is no help.

9. Back to propositions

Let us take sentential complements of the form 'that *p*' to contribute a proposition to the second argument position in '*s* in *L* means *y*' (for simplicity I drop relativization to context). What kind of singular term is 'that *p*'? We seem to have three options.

1. it is a description (construed as a quantifier)
2. it is a referring term that merely introduces an object into the proposition
3. it is a term that refers via a Fregean mode of presentation

My claim is this: no matter which of these options we take, it turns out that a true sentence of the form '*x* in *L* means *y*' enables us to understand '*x*' only if we can construct from '*y*' a metalanguage sentence that we understand and understand to be the same in meaning as '*x*'.[2]

Option 1. 'that *p*' is a description. What is the form of the description? The most natural thing to say is that, as the sentence '*p*' expresses the proposition, the description is 'the proposition expressed by "*p*" in English'. Alternatively, we could treat it as a context sensitive self-referential description: 'the proposition the speaker of this token "*p*" expresses with it'.

φ means in *L* the proposition expressed by '*p*' in English
φ means in *L* the proposition the speaker of this token '*p*' expresses with it.

It is clear, however, that this would not put someone in a position to understand the object language sentence unless he understand the mentioned sentence or the mentioned sentence as used by the speaker. But what if it were some non-metalinguistic description? Any proposal along these lines will be subject to the objections to option 3.

Option 2. 'that *p*' contributes only its referent to what is said. We may still compatibly with this take the referent to be determined by a description. We could think of it functioning like Kaplan's dthat(the *F*).

φ means in *L* dthat(the proposition expressed by '*p*' in English)

[2] I give a parallel argument to show that reference to propositions is no help in understanding how attitude attributions help us to understand what people think in (Ludwig 2014).

But even so, what is said clearly doesn't put one in a position to understand φ independently of understanding '*p*'. For the same thing could be said using a directly referring term. Let us name the proposition expressed 'Bob'. We say the same thing then in saying:

φ means in *L* Bob

But clearly someone could understand what is said by this without understanding φ.

Option 3. Frege held that '*p*' in the context following 'means that' has an indirect sense that is a mode of presentation of its customary sense, i.e., the proposition expressed by '*p*'. Of course, one mode of presentation might simply be 'the proposition expressed by '*p*' in English'. But it is clear already that this will not give us understanding of the object language sentence independently of understanding '*p*' in English. What other mode of presentation might we appeal to? I venture to say that the only idea anyone has of this is given by the description of the role it is to play. But whatever it is, we know that it has to meet a certain constraint, a constraint which I think cannot be met. The constraint is that the mode of presentation must be such that one cannot present to oneself the proposition via that mode of presentation without grasping it. For otherwise, one could grasp the proposition expressed by 'φ means in *L* that *p*' without understanding φ. Grasping a proposition occurrently, as is required here, entails entertaining it. So what we require is a mode of presentation attached to 'that *p*' of which it is constitutive that one entertain its object. But the mode of presentation is distinct from what it presents, and so entertaining it is not ipso facto to entertain its object. So for no mode of presentation of an entertainable object could grasp of the mode of presentation suffice for entertaining its object. (See (Ludwig 2014) for further discussion, and see especially section 5.)[3]

But there is an additional problem with the appeal to a non-metalinguistic mode of presentation, namely, that it would make the fact that '*p*' appears in 'that *p*' an accident of spelling. For whatever the non-metalinguistic sense that is to do the job could be attached to any arbitrary term, like 'Bob', and function in exactly the same way. But it is obvious that using a sentence we understand in the complement is crucial to the way these sentences inform us about the meaning of the sentences they are about. It is not just an accident of spelling. So even if we could make sense of a mode of presentation of a proposition that

[3] The new cognitivist accounts of propositions championed by Soames (King, Soames, and Speaks 2014; Soames 2010, 2015) and Hanks (Hanks 2015) might be thought to provide a way around this difficulty. The basic idea is that propositions are cognitive act types, of which the basic act type is that of predicating a property of an object. The relevance of this to attitude attributions is that in determining the cognitive act type to be associated with, e.g., 'Russell believed that mathematics was reducible to logic', we consider what sequence of act types are instantiated when someone understands that sentence. Since that involves understanding 'mathematics was reducible to logic', the act type is different from that involved in grasping 'Russell believed Logicism', and it explains why in grasping the former one has to entertain the proposition the 'that-clause' refers to, while that is not so with respect to the latter. But while that suffices to distinguish the propositions and explain why the former involves entertaining the proposition expressed by the sentence, it doesn't show that understanding the sentence in the complement is not essential to the way it functions in the language. It does not provide an account of a Fregean sense grasp of which suffices to entertain its object. It is in fact fully compatible with the view that '*s* means that *p*' conveys to us what *s* means by way of our understanding '*p*' and understanding it to translate *s*.

guaranteed occurrent grasp of it, this would be a fatal objection to a non-metalinguistic Fregean account: it could avoid the problem only by ignoring a central feature of the mechanism by which sentential complements do their work for us.

It is clear that as a matter of fact the way we are clued into the meaning of an object language sentence via an M-theorem for it is by way of our understanding the metalanguage sentence and the requirement that the complement sentence be the same in meaning as the object language sentence, relative to context. And it seems clear that anything you put in the second argument place in that construction will not put you in a position to understand the object language sentence except insofar as it at least codes for such a sentence. The difficulty is that understanding a sentence is not at bottom a matter of standing in a relation to an abstract object and associating it with a sentence. It is a matter of knowing how to use it for certain purposes. To convey what a sentence means we can rely on antecedent understanding of terms appropriately related in meaning, or we can explain the purpose of the sentence in the kind of activity that is central to its linguistic meaning. The second of these, however, will not take the form of relating a sentence to any object.

One could insist on a very abstract reading of mode of presentation, and allow that the mode of presentation of a proposition may simply consist in our understanding a sentence and thinking of the proposition as the one we are grasping in understanding the sentence. But this would just show that the work that sentences in the complements of M-sentences do depends essentially on our understanding them and knowing the claim is that the sentence which is its subject is the same in meaning as it.

This point extends to the various ways we have of trying to indicate the meanings of subsentential expressions by assigning them properties and relations and functions. To make meaning scrutable by assignment of an object to an expression, we must use an expression that at least codes for an antecedently understood expression which is understood to interpret the expression whose meaning we are given (with the except of referring terms whose referents are fully given by a rule relative to contextual parameters). This is transparent in, for example, the claim that ‘rouge’ in French means the property of being red, or ‘aime’ in French means the relation of loving.

The conclusions to reach are the following:

> 1. The assignment of entities to expressions in pursuit of a meaning theory for a natural language is neither necessary nor sufficient. It is not necessary because the same goals, so far as possible, can be achieved by the method of truth-theoretic semantics. It is not sufficient because assigning the correct entities alone does not put us in a position to understand object language expressions. We must assign entities using terms that code for expressions we understand and understand to interpret object language expressions to which meaning entities are being assigned.
>
> 2. A theory of meaning designed to issue in theorems of the form ‘φ means in *L* that *p*’ does its work inevitably in part by way of showing us something about what object language expressions mean by way of our antecedent understanding of metalanguage terms understood to interpret them. There is, thus, a limit to the depth of the illumination of meaning in the object language we can expect. We have

> nominally specified a body of propositional knowledge sufficient to understand any sentence in a language, but on closer examination, antecedent understanding of sentences understood to be alike in meaning with object language sentences plays a crucial role. To get greater illumination, we have to move to a different form of account. (This is obviously connected with the dispute between Davidson and Dummett (Dummett 1975, 1976) over whether a modest theory of meaning is adequate to our philosophical purposes.)

9. The theory of meaning

Let me conclude by returning to the question of the relation between the project of giving a meaning theory for a language and of giving a theory of meaning. The ambition of a meaning theory is to state knowledge that suffices for one to understand any utterance of a sentence in the language. Ultimately the theory should issue in explicit statements of what sentences in the language mean, as a way of expressing the sort of knowledge we want it to give us, whether we pursue the project directly or indirectly, that is to say, it should issue in M-thereoms.[4] This is not a full theory of meaning, but if successful it would make an important contribution. It would provide a body of propositional knowledge independent of knowledge of understanding any particular language that would determine the meaning facts about a particular language in a way that allows understanding of the language. However, surprisingly, the very ambition to provide an explicit M-theorem for every sentence shows that what insight the theory gives relies upon prior understanding of a language together with knowledge that a sentence one understands is the same in meaning as the sentence whose meaning one wants to grasp. If this is right, then no theory of this sort can achieve its ambition.

Davidson hoped to bridge the gap between a meaning theory and a theory of meaning, as I have noted, by an account of how a truth theory could be confirmed for a speaker on the basis of evidence that did not presuppose anything about a speaker's words or any detailed knowledge of his attitudes, that is, from the standpoint of the radical interpreter (Davidson 1973b). This is a way of connecting up the structure articulated in a truth theory with the use of words by speakers of the language for which it is theory in a way, it was hoped, that would guarantee its canonical theorems were interpretive. The general idea was that we would gain insight into the content of the relevant concepts by seeing how a theory sufficient for interpretation could be confirmed on the basis of evidence that did not presuppose application of the concepts of the theory. Davidson's fundamental assumption about radical interpretation was that a correct meaning theory could be recovered ultimately from purely behavior evidence and what we can know about agents and speakers a priori. I won't argue for it here, but I think this assumption is mistaken (Lepore and Ludwig 2005).

This still leaves us with the task of relating language—words and expressions, and ultimately sentences—to the uses to which they are put and the purposes for which they are designed. What I believe this requires is a theory of communicative institutions and the forms of collective agency that underlie them (Jankovic 2014a, 2014b). If we cannot hope for a reduction of meaning to behavioral responses to the environment, then we must

[4] These remarks therefore apply as well to Ray's account discussed in note 2.

relate meaning to the attitudes people have in using words and expressions in communicative contexts. This will inevitably rest on an antecedent understanding of how we are able to represent the world in thought. But it holds out the hope of breaking out of the circle of linguistic concepts, which the fixation on generating M-theorems in the theory of meaning does not.

References

Boisvert, Daniel, and Kirk Ludwig. 2006. Semantics for Nondeclaratives. In *The Oxford Handbook of the Philosophy of Language*, edited by B. Smith and E. Lepore. Oxford: Oxford University Press.

Davidson, Donald. 1965. Theories of Meaning and Learnable Languages. In *Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science.*, edited by Y. Bar-Hillel. Amsterdam: North Holland Publishing Co.

———. 1967. Truth and Meaning. *Synthese* 17:304-323.

———. 1968. On Saying That. *Synthese* 19:130-146.

———. 1970. Semantics for Natural Languages. In *Linguaggi nella Societa e nella Tecnica*. Milan: Comunita.

———. 1973a. In Defence of Convention T. In *Truth, Syntax and Modality*, edited by H. Leblanc. Dordretch: North-Holland Publishing Company.

———. 1973b. Radical Interpretation. *Dialectica* 27:314-328.

———. 1974. Belief and the Basis of Meaning. *Synthese* 27:309-323.

———. 1975. Thought and Talk. In *Mind and Language*, edited by S. Guttenplan. Oxford: Oxford University Press.

———. 1976. Reply to Foster. In *Truth and Meaning: Essays in Semantics*, edited by G. Evans and J. McDowell. Oxford: Oxford University Press.

———. 1979. Moods and Performances. In *Meaning and Use*, edited by A. Margalit. Dordrecht: D. Reidel.

———. 2001. Semantics for Natural Languages. In *Inquiries into Truth and Interpretation*. New York: Clarendon Press. Original edition, 1970.

Dummett, Michael. 1975. What is a theory of meaning? In *Mind and Language*, edited by S. Guttenplan. Oxford: Oxford University Press.

———. 1976. What is a Theory of Meaning? (II). In *Truth and Meaning: Essays in Semantics*, edited by G. Evans and J. McDowell. Oxford: Oxford University Press.

Hanks, Peter. 2015. *Propositional content*. New product edition. ed, *Context and content*. New York, NY: Oxford University Press.

Hoeltje, Miguel. 2013. Lepore and Ludwig on 'explicit meaning theories'. *Philosophical Studies* 165 (3):831-839.

———. 2016. 'Meaning and Truth' and 'Truth and Meaning'. *Dialectica* 70 (2):201-215.

Jankovic, Marija. 2014a. Communication and Shared Intention. *Philosophical Studies* 169:489-508.

———. 2014b. Conventional Meaning. Dissertation, Philosophy, Indiana University.

King, Jeffrey C., Scott Soames, and Jeffrey Speaks. 2014. *New thinking about propositions*. First edition. ed. Oxford: New York Oxford University Press.

Lepore, Ernest, and Kirk Ludwig. 2005. *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Oxford University Press.

———. 2007a. *Donald Davidson: Truth-theoretic Semantics*. New York: Oxford University Press.

———. 2007b. Radical misinterpretation: A reply to Stoutland. *International Journal of Philosophical Studies* 15 (4):557-585.

———. 2011. Truth and Meaning Redux. *Philosophical Studies* 154:251-277.

Ludwig, Kirk. 2003. The Truth about Moods. In *Concepts of Meaning: Framing an Integrated Theory of Linguistic Behavior*, edited by G. Preyer, G. Peter and M. Ulkan. Dordrecht: Kluwer Academic Publishers.

———. 2014. Propositions and higher-order attitude attributions. *Canadian Journal of Philosophy* 43 (5-6):741-765.

———. 2015. Was Davidson's Project a Carnapian Explication of Meaning? *The Journal of the History of Analytic Philosophy* 4 (3):1-55.

Ray, Greg. 2014. Meaning and Truth. *Mind* 123 (489):79-100.

Soames, Scott. 2008. Truth and Meaning: In Perspective. *Truth and Its Deformities: Midwest Studies in Philosophy* 32:1-19.

———. 2010. *What is meaning?*, *Soochow University lectures in philosophy*. Princeton, N.J.: Princeton University Press.

———. 2015. *Rethinking language, mind, and meaning*, *The Carl G Hempel lecture series*. Princeton ; Oxford: Princeton University Press.

Wiggins, David. 1980. 'Most' and 'All': Some Comments On a Familiar Programme. In *Reference, Truth and Reality: essays on the philosophy of language*, edited by M. Platts. London: Routldge and Kegan Paul.