Davidsonian Theories of Meaning

What is a theory of meaning?

Not, for Davidson, a theoretical account of the notion of meaning, but a theory of the meanings of sentences of a particular language (say, French or English). A theory of meaning for a language is an empirical theory. The philosophical work comes in describing the form a theory of meaning for a language will take and the nature of the evidence that will count as confirming it. When we sort out these issues, Davidson thinks, there won't be any further questions about what "meaning" is, or what "meanings" are. (In fact, a Davidsonian theory of meaning gets by without mentioning "meanings" at all!)

What is the aim of a theory of meaning?

To state "something knowledge of which would suffice for interpreting utterances of speakers of the language to which it applies" (*Inquiries into Truth and Interpretation*, 171; see also "Radical Interpretation," 125). To interpret a speaker is to say what is said by her utterances, e.g. to redescribe Sally's utterance of the words 'Es schneit' as an act of saying that it is snowing ("Belief and the Basis of Meaning," 141).

What form should a theory of meaning take?

The theory must give the meanings of infinitely many sentences. But this knowledge must be finitely specifiable, if languages are to be learnable. So a theory of meaning can't consist in an infinite list of sentences and their meanings. The meanings must be generated recursively. Accordingly, Davidson proposes that a theory of meaning take the form of a Tarskian truth definition for a language:

> There is no need to suppress, of course, the obvious connection between a definition of truth of the kind Tarski has shown how to construct, and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. To know the semantic concept of truth for a language is to know what it is for a sentence—any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language. (*Inquiries*, 24)

What will the clauses of a Tarskian truth definition look like?

Since the definition is recursive, there will be two kinds of clauses, *base clauses* and *recursive clauses*. In the simplified truth definition we constructed in the handout on Tarskian truth definitions, the base clauses tell you the truth conditions of simple sentences (e.g., 'Joe is fat' is true-in-English iff Joe is fat), and the recursive clauses tell you how to figure out the truth conditions of complex sentences in terms of the truth conditions of their parts (e.g., 'A and B' is true-in-English iff 'A' is true-in-English and 'B' is true-in- English).

For more complex languages (e.g., those with quantifiers like 'someone' or 'the'), this won't work. Tarski saw that in order to give a truth definition for such a language, one must talk of *satisfaction* conditions instead of truth conditions. Roughly, to state the satisfaction conditions of a predicate (or open sentence) is to say which objects it is true of. This complication is important if you're actually constructing a truth definition. If you're just testing one, you can ignore the "machinery" and just look at the T-sentences it churns out. ("Radical Interpretation," 153-4; on T-sentences, see below.)

How can Davidson handle indexicals, demonstratives, and tense?

Tarski's formalized languages contained no indexicals ('I', 'now'), demonstratives ('this'), or tense. In order to accommodate these devices in a truth definition, one must relativize truth to (at least) a speaker and a time. So, instead of

'I am tired now' is true-in-English iff I am tired now,

which clearly wouldn't work, the truth definition spits out

'I am tired now' is true-in-English as spoken by speaker S at time t iff S is tired at t.

Why a truth definition, and not a (recursive) method for translation?

"When interpretation is our aim, a method of translation deals with a wrong topic, a relation between two languages, where what is wanted is an interpretation of one (in another, of course, but that goes without saying since any theory is in some language)" ("Radical Interpretation," 129). Pierre, a monolingual Frenchman, could know that 'schnee ist weiss' is the German translation of the English 'snow is white' without knowing what either means. Also, translation is plainly inappropriate in interpreting utterances containing indexicals and demonstratives. You can translate an utterance of 'I am tired' into French without knowing who the speaker is, but you can't interpret it (understand what it says) unless you know who the speaker is.

How is a theory of meaning tested?

A theory of meaning is tested, like any empirical theory, through its consequences. The consequences of a theory of meaning are the (infinitely many) T-sentences it implies.

What are T-sentences?

T-sentences (or T-biconditionals) are sentences of the form 'S is true iff p,' where S is a name or description of a sentence and p is a sentence. Examples:

1. 'Sam loves Rhonda and Rhonda loves Sam' is true-in-English iff Sam loves Rhonda and Rhonda loves Sam.
2. 'Schnee ist weiss' is true-in-German iff snow is white.
3. 'Snow is white' is true-in-English iff leopards are animals.

Note that 'iff' ('if and only if') here is the material biconditional: 'A iff B' is equivalent to '(A and B) or (not-A and not-B)' (there need not be any relevant connection between A and B). The left side of a T-biconditional *mentions* a sentence in the object language (the language for which the theory of meaning is being given). The right side of the biconditional *uses* a sentence in the metalanguage (the language in which the theory of meaning is being framed). As example (3) demonstrates, the sentence on the right side need not be a translation into the metalanguage of the sentence mentioned on the left side. Davidson says:

> ...an acceptable theory of truth must entail, for every sentence s of the object language, a sentence of the form: s is true if and only if p, where 'p' is replaced by any sentence that is true if and only if s is. Given this formulation, the theory is tested by evidence that T-sentences are simply true; we have given up the idea that we must also tell whether what replaces 'p' translates s. (RI, 134)
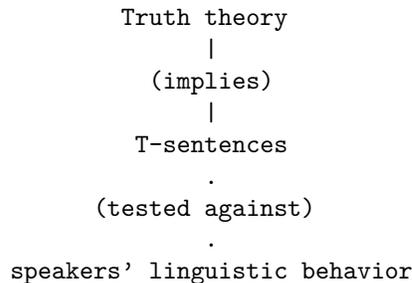
Davidson compared with Tarski

Tarski said that a truth definition for a language could be counted correct only if it implied every sentence of the form 'S is true-in-L iff p' where p is a translation into the metalanguage of the object-language sentence S. (This was his "Convention T.") Thus, if you know what sentences of the object language mean (that is, how they are to be translated), you can evaluate a definition of truth for that language. Davidson reverses Tarski's direction: he claims that if you know which T-sentences are true, you can evaluate a theory of meaning for a language. "...assuming 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" (RI, 134; cf. BBM, 150).

What counts as evidence for the truth of the T-sentences?

We can't directly see that a T-sentence is true. We have to infer its truth from what we do see: the speaker's use of the sentence. Davidson assumes that we can start with some pretty good guesses about which sentences the speaker holds true. Holding a sentence true "...is an attitude an interpreter may plausibly be taken to be able to identify before he can interpret, since he may know that a person intends to express a truth in uttering a sentence without having any idea what truth" (RI, 135) So, "...the evidence available is just that speakers of the language to be interpreted hold various sentences to be true at certain times and under specified circumstances" (RI, 135). (Note: this is not, strictly speaking, behavioral evidence: holding-true is a mental state, not a behavior, though it more directly manifested in behavior than other mental states.)

Review

We've got a Tarski-style truth theory that implies infinitely many T-sentences, i.e., sentences of the form 'S is true-in-L (as uttered by S at t) iff p'. Now we want to know whether this theory is a correct theory of meaning for L. Davidson says that the T-sentences are the point at which the theory is tested against empirical reality. (The machinery doesn't matter, as long as it spits out the right T-sentences.) But what is it, exactly, against which we test the T-sentences? Answer: speakers' use of sentences of the language. So we've got a picture like this:

```
            Truth theory
                 |
             (implies)
                 |
            T-sentences
                 .
          (tested against)
                 .
     speakers' linguistic behavior
```

How do we move from this evidence to judgments of the truth of the T-sentences?

Here's where things get interesting. We can't just assume that every sentence the speaker holds true is true. Speakers have false beliefs; they sometimes take false sentences to be true.
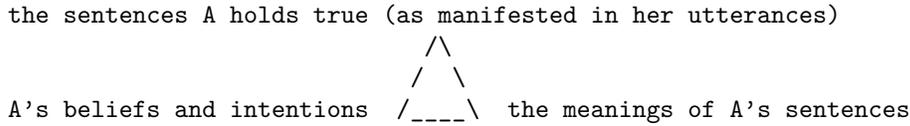
> A speaker who holds a sentence to be true on an occasion does so in part because of what he means, or would mean, by an utterance of that sentence, and in part because of what he believes. If all we have to go on is the fact of honest utterance, we cannot infer the belief without knowing the meaning, and have no chance of inferring the meaning without the belief. ("Belief and the Basis of Meaning," 142)

Example: Sally sincerely utters 'George is blezzy.' If we knew what 'blezzy' meant (say, flirtatious), then we could infer that Sally believes that George is flirtatious. If we knew that Sally believed (and intended to say) that George is flirtatious, then we could infer that 'blezzy' means flirtatious. But how can we get started if we know neither?

Can't we know what someone believes or intends without knowing her language? Davidson assumes that in general we won't be able to figure out someone's beliefs and intentions without being able to interpret her language: "...making detailed sense of a person's intentions and beliefs cannot be independent of making sense of his utterances" (BBM, 144). "I conclude that in interpreting utterances from scratch—in *radical* interpretation—we must somehow deliver simultaneously a theory of belief and a theory of meaning. How is this possible?" (BBM, 144)
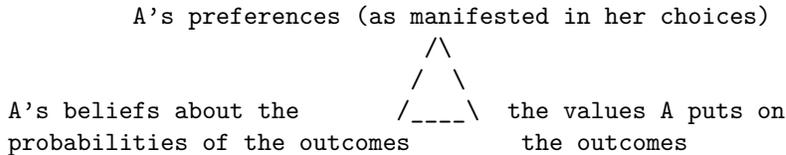
The triangle

As Davidson says, linguistic behavior is a vector of belief and meaning. Given two corners of this triangle, we could "solve for" the third. But how can we make any progress if we only have the top one? This is the problem of radical interpretation.

```
  the sentences A holds true (as manifested in her utterances)
                              /\
                             /  \
  A's beliefs and intentions /____\  the meanings of A's sentences
```

An analogy from decision theory

In "Belief and the Basis of Meaning" (BBM), Davidson points out that this problem is analogous to a problem that arises in decision theory, except here the triangle is:

```
          A's preferences (as manifested in her choices)
                              /\
                             /  \
  A's beliefs about the     /____\  the values A puts on
  probabilities of the outcomes      the outcomes
```

Suppose Bert is indifferent between getting $5 (guaranteed) or taking a bet on getting $20 if the coin lands Heads and $0 if it lands Tails. There are three possibilities:

1. Bert is being irrational.
2. Bert doesn't value $20 four times as much as he values $5; he only values $20 twice as much as he values $5.
3. Bert believes that the coin is rigged, so that it is more likely to land Tails than it is to land Heads.

If we don't rule out (1), we can't get anywhere. So we assume that Bert is acting rationally. That leaves (2) and (3): how do we choose between them? We can't, as long as we confine our attention to this single choice. But if we look at other possible choices, we can. Offer Bert the opposite choice: $5 guaranteed, or $20 if the (same) coin lands Tails and $0 if it lands Heads. If Bert believes the coin is rigged towards Tails (3), he'll take the bet. But if he simply doesn't value $20 four times as much as he values $5 (2), he'll remain indifferent between taking the bet and taking the $5.

Morals of the story:

1. We can't explain Bert's choices at all unless we assume that he's rational. (At least we can't give this kind of explanation: an explanation of his reasons for choosing as he does. We might be able to explain why he's acting strangely by noting, say, that he has a brain lesion. But that's a different kind of explanation.)

2. We can't test a hypothesis that is solely about why Bert makes this choice. We can only test a theory (an ascription of subjective probabilities and subjective values to Bert) that makes predictions about many possible choices: "…to explain (i.e. interpret) a particular choice or preference, we observe other choices or preferences; these will support a theory on the basis of which the original choice or preference can be explained" (BBM, 146).

How does this apply to interpreting utterances?

1. We must apply a principle of charity in interpretation. We must assume that the speaker is rational and that her beliefs are, for the most part, true. Thus, for example, if Sally is standing right in front of George and sincerely asserts 'George is blezzy,' then we should not interpret 'blezzy' as meaning tall if George is plainly short. If Jim says 'Joan is swimming and Joan is blip swimming,' we should not interpret 'blip' as meaning not. This doesn't mean that we can't ascribe false beliefs to speakers, just that we should do so only when necessary, and only when it makes sense for the speakers to have false beliefs in their circumstances. Note:

   > The methodological advice to interpret in a way that optimizes agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances and other behavior of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything. (RI, 137)

2. "...to interpret a particular utterance it is necessary to construct a comprehensive theory for the interpretation of a potential infinity of utterances" (BBM, 148). For example, we must look at Sally's uses of other sentences involving 'blezzy', like 'Jim is not blezzy', 'a blezzy person would not do that', etc. Individual T-sentences cannot be tested; a theory of meaning must be tested as a whole.

Do all true T-sentences give the meanings of the sentences mentioned on their left sides?

No. To see this, note that the T-sentence

'snow is white' is true-in-English iff grass is green

is perfectly true. (Remember, 'iff' is the material biconditional. So 'A iff B' is true if A and B are both true or both false, no matter what they say.) The problem, as Davidson points out, is that "the T-sentence does fix the truth value relative to certain conditions, but it does not say the object language sentence is true because the conditions hold" (RI, 138, cf. BBM, 150).

Under what conditions does a true T-sentence give the meaning of a sentence?

Davidson's view is that a true T-sentence gives the meaning of the sentence mentioned on its left-hand side when it can be derived (in a canonical way) from a recursive theory of truth for a language that optimally fits all of the evidence. This should rule out bogus T-sentences like " 'snow is white' is true-in-English iff grass is green," because the recursive theory that implies this will also imply " 'snow falls from the sky' is true-in-English iff grass falls from the sky," which is false (and will be revealed as false through the process of radical interpretation). (See RI, 139.)

The theory of meaning as part of rational psychology.

The key Davidsonian idea is that the theory of meaning is part of a larger enterprise of "rational psychology," the project of making sense of what people do—that is, of understanding their reasons for what they do. We explain people's actions by adverting to their beliefs and desires. Why did Joe open the cupboard? Because he wanted a glass and he believed that the glasses are stored in that cupboard. Davidson holds that the theory of meaning is interdependent with belief-desire psychology:

- Ascriptions of beliefs and desires depend on the interpretation of language: Theories about someone's beliefs and desires cannot (in general) be tested unless we know how to interpret her speech.
- The interpretation of language depends on ascriptions of beliefs and desires: Theories of meaning cannot be tested unless we make assumptions about speakers' beliefs (and desires).

On Davidson's view, neither the theory of meaning nor belief-desire psychology is prior to the other. They're part of a single package that can only be tested as a unit. "Broadly stated, my theme is that we should think of meanings and beliefs as interrelated constructs of a single theory just as we already view subjective values and probabilities as interrelated constructs of decision theory" (BBM, 146).

Is the interdependence merely epistemic?

So far, all we've said is that a theory of meaning and an ascription of beliefs and desires can only be tested in conjunction, so that knowledge of one presupposes knowledge of the other. But Davidson doesn't think that the interdependence of interpretation and rational psychology is merely epistemic, merely a matter of our knowledge. "The semantic features of language are public features. What no one can, in the nature of the case, figure out from the totality of the relevant evidence cannot be part of meaning" (Inquiries, 235). On Davidson's view, there can't be any facts about meaning other than what can be discerned through radical interpretation. Similarly, there can't be any facts about one's beliefs and desires other than what can be discerned in the holistic process of "making sense of you." In particular, beliefs and desires are not something the natural sciences can tell us anything about (see the last paragraph of BBM, which gets elaborated in "Mental Events").

Davidson compared and contrasted with Quine.

Davidson's view is clearly quite close to Quine's (see BBM 148-9):

- Both are behaviorists, in the sense that contrasts with "psychologism": they think that only publicly available evidence is relevant to the assessment of a translation manual or theory of meaning. This is reflected in their theoretical use of the notion of radical translation or radical interpretation.
- Both are holists: they hold that questions of interpretation or translation can't be decided piecemeal for individual sentences.
- Both reject any principled distinction between questions of meaning and questions of fact.
- Both deny that intentionality is a topic for the natural sciences (BBM, 154).
- Both accept that there will be some indeterminacy in translation or interpretation (RI, 139, BBM, 153-4). Davidson suggests that this indeterminacy need be no more worrisome than the fact that we might model someone's preferences by assigning the numbers 1, 2, and 3 to three outcomes or by assigning the numbers 2, 4, and 6 (BBM, 147, 153- 4).

However, there are also some differences (see RI 129 n. 3):

- Where Quine talks of *translation*, Davidson talks of *interpretation*. Knowing a translation manual, which just pairs sentences with sentences, would not suffice for understanding a language (one might not understand either of the paired sentences). But knowing a Davidsonian theory of meaning, which gives truth conditions for sentences, would (plausibly) suffice for understanding a language.
- They differ on the nature of the objective evidence on which an interpreter can base translation. For Davidson, the evidence includes information about what sentences speakers hold true in what environmental conditions (e.g., the fact that the natives tend to hold true 'gop gop' when it's raining in the near vicinity). Quine starts with a more austere basis: stimulus meanings.
- For Davidson, the principle of charity is constitutive of correct interpretation. For Quine, charity is merely a pragmatic maxim we use in constructing translation manuals, but a more charitable translation manual (one that makes more of the natives' claims come out true) is not thereby more correct. (One upshot is that Davidson will not have to admit as much indeterminacy as Quine.)

Davidson contrasted with Grice and Searle

For Davidson, it doesn't make sense to analyze meaning in terms of beliefs and intentions, because all of these notions are conceptually interdependent. That is, we can't make sense of any of them apart from

the others. So there's not going to be any "one way arrow" between beliefs and intentions and meaning. (See RI, 127, BBM, 144.)

Note that Grice and Searle would agree with Davidson that we can't (in general) come to know what someone believes or intends without being able to interpret her speech, but they would say that this is merely an epistemological point, and that it doesn't bear on the question of what it is for someone to have a particular belief or intention. It is Davidson's behaviorism that allows him to make the move from the epistemology to the metaphysics.

Note also that whereas the Gricean and Searlean stories make meaning ultimately a matter of the speaker's intentions, Davidson's account of meaning involves essential reference to an (actual or hypothetical) interpreter.

Davidson compared with Putnam and Burge

Like Putnam and Burge, Davidson is an externalist. Since the meanings of our words depends on what would be the best way for someone to interpret us, and since that depends on what is going on outside our heads, meaning can depend on what is going on outside our heads. The best way to interpret a brain in a vat (assuming it has been in the vat a long time) is to take its beliefs, desires, and words to be about the states of the computer that is giving it stimuli. But Davidson differs from Burge in not giving the weight he does to the "public language." This comes out clearly in "A Nice Derangement of Epitaphs."