Critical notice of Davidson's *Inquiries into Truth and Interpretation*

Jonathan Bennett

from: *Mind* 94 (1985), pp. 601–626.
concerning: Donald Davidson, *Inquiries into Truth and Interpretation* (Clarendon Press, 1984.)

This book holds eighteen papers about language, produced by Donald Davidson at a steady rate through the eighteen years from 1965 onwards. A famous theory about meaning is the chief topic of papers 1, 2, 4, 8–12, 15, and looms large in 13 and 14 also. I will thread most of this review on that theory, with passing mentions also of papers 3 and 5–7. Before embarking on all that, I shall say a little about the final three papers.

The final three papers

In 'The Inscrutability of Reference' (paper 16), Davidson agrees with Quine that what we refer to by our singular terms and predicates is unavoidably indeterminate, but he doesn't agree that ontology must be relativized. We must decide, perhaps arbitrarily, whether the speaker we are studying uses 'rabbit' to refer to rabbits or rather to things that are R to rabbits, and that is a decision about how to interpret his sentences by means of sentences of ours. That is relativity of a sort, it is the most that can be extracted from Quine's premise, and it should not be called 'ontological relativity'. Davidson objects to that phrase because it suggests this (the formulation is mine):

> Someone who uses the word 'rabbit' in such a sentence as 'There's a rabbit over there' refers indeterminately; but he could instead say something of the form 'There's a rabbit$_L$ over there', relativizing his word 'rabbit' to language L and thereby making it perfectly determinate. Analogously, 'The packing of the Supreme Court was useful' has an indeterminacy which is absent from 'The packing of the Supreme Court was useful to FDR'.

Davidson is clearly right to reject that, and I imagine Quine would reject it too. That leaves the question of what Quine does mean when he speaks of 'ontological relativity', and I join Davidson in being uncertain about this. This is a densely argued and, in my opinion, highly successful paper.

It has often been maintained **i** that what a metaphor valuably achieves is to express some true cognitive content, and **ii** that it does this by using words in metaphorical as distinct from literal meanings. In 'What Metaphors Mean' (paper 17), Davidson denies not just **ii** but even the weaker **i**, primarily on the grounds that if **i** were true then it should not be impossible, as apparently it is, to replace any given metaphor by an equivalent non-metaphor. Metaphors are all false, Davidson says, but they can valuably get us to see things in new lights, and so on. This is a stimulating, knock-about piece, which should be read along with counter-attacks by Black and Goodman (see p. xii).

The main content of 'Communication and Convention' (18) is a discussion of David Lewis's theory that a convention is a certain kind of behavioural regularity that solves a recurring co-ordination problem—the problem, in the case of language, being to bring it about that speaker and hearer pair sounds with meanings in the same way. Davidson thinks that the 'most important feature of Lewis's analysis of convention', namely its use of the concept of regularity, is its Achilles' heel: 'The only candidate for recurrence we have is the interpretation of sound patterns: speaker and hearer must repeatedly. . . interpret relevantly similar sound patterns of the speaker in the same way (or ways related by rules that can be made explicit in advance)' (p. 278), and Davidson doubts whether any such idea is of much use in explaining and describing communication.

His doubts stem from the fact that a hearer often interprets a speaker in ad hoc ways, making adjustments to his prior expectations in the light of conversational reality.

Everyday communication does indeed include a lot of this sort of thing. We interpret one another's sentences pretty well, despite breaks and stumbles and errors of all kinds. If my 'advance theory' about someone includes the hypothesis that he uses the word 'connive' in its original sense of 'wink at' or 'pretend not to notice', and then I hear him say 'N is a conniving swine' when there is no question of N's noticing or ignoring bad behaviour by someone else, I will smoothly revise my advance theory and interpret the speaker as meaning that N is conspiratorial and manipulative.

According to Davidson, such ad hoc adjustments are not applications of any convention, because they do not involve applying rules that both speaker and hearer had internalized in advance ('Unless [speaker and hearer] coincide in advance, the concepts of regularity and convention have no definite purchase'). Indeed, they do carry into the situation intellectual possessions that help in the adjustment, but nothing that could fairly be called rules: 'The speaker must have some idea of how the hearer is apt to make use of the relevant clues; and the hearer must know a great deal about what to expect. But such general knowledge is hard to reduce to rules, much less conventions or practices.'

Well, we cannot write the rules out, but unlike Davidson I see evidence that they exist and guide us when as hearers we make emergency repairs in our interpretative theories. In general, two hearers will uncollusively deal in the same way with a speaker's deviations from their advance theory about him, which strongly suggests that they are unconsciously applying some shared rules for the handling of such deviations. There is indeed a small literature, not mentioned by Davidson, about what the rules are.

Still, even if he is wrong to stop at 'regularity', Davidson could properly have dug in his heels one step further back, at 'convention'; though the argument for this could hardly have occurred to him unless he stopped bustling through Lewis's work and started listening, patiently and attentively, to what Lewis has to say. The argument goes as follows.

In the event, I interpret the speaker as using 'connive' to mean conspire, and that's what he expected me to do, which is why he used the word with that meaning. That sounds a little like a convention, but really it is not. If it were, the speaker's expectation and my interpretation would rest on shared true beliefs about one another's handlings of the word 'connive'; whereas in fact he was going by a false belief about how I regularly understand 'connive', and I went by a belief he doesn't share, namely that he has made some mistake, probably the mistake of thinking that I take 'connive' to mean conspire.

In a genuine convention, according to Lewis's analysis, everything is open and above-board, and everyone is on the

same epistemic level. Davidson is right that our adjustments to defective speech involve our often not being on a level, and involve some of us in trains of thought to which others are not privy. In my *Linguistic Behaviour* I describe possible languages—those of creatures I call Condescenders—where communication is *never* level and open, though of course I don't think that human languages are like that.

Still, Lewis's concept of convention fits a lot of what goes on, and its power and depth are manifested in how it helps our understanding of the linguistic transactions to which it *doesn't* apply. In other ways, too, it is a superb achievement. It provides a central core onto which the concepts of regularity, rule, and norm can be helpfully fixed. It explains why meaning conventions matter although they are in a way arbitrary, this being not a paradox but of their essence. It frees us from naïve contrasts between intention and convention, by showing how the two are interrelated.

Davidson does not acknowledge this, and presumably has not seen it. Lewis seems not to be one of those few philosophers whose work he will attend to with care. It is a pity there are not more. In particular, I believe that Davidson in his theory of meaning or interpretation tries to build bricks without straw, and that he might have seen this if he had attended to the work of Grice and his followers and of Lewis.

(A recent paper of Davidson's, not included in this volume, should be parenthetically mentioned at this point. It questions the role that is played in linguistic communication by knowledge of meanings that is 'systematic, shared and prepared'. The target now is not only the concept of convention but a whole range of 'standard descriptions of linguistic competence (including descriptions for which I am responsible)'.[1]

Although things could conceivably have been otherwise, Davidson is right in saying that in actual human conversation we are opportunistically ingenious in understanding various kinds of defective speech, differences of idiom, new proper names, and so on. Given how we speak, we would not thrive as speakers or hearers if we were armed only with what is systematic, shared and prepared: 'The general framework or theory, whatever it is, may be a key ingredient in what is needed for interpretation, but it can't be all that is needed since it fails to provide the interpretation of particular words and sentences as uttered by a particular speaker' (p. 23). In Davidson's hands, this modest and not unfamiliar point is made to look radical and iconoclastic, the illusion being created by repeated blurrings of the distinction between what suffices for communication to succeed and what plays an important part in successful communication, as when Davidson criticizes the suggestion that x is 'essential' for communication on the grounds that x is not 'adequate' for communication (pp. 22f.).

Here is his strongest statement of his conclusion: 'What interpreter and speaker share, to the extent that communication succeeds, is not learned and so is not a language governed by rules or conventions known to speaker and interpreter in advance' (p. 24). The drastically unclear phrase 'share, to the extent that communication succeeds' sprawls across the difference between **(i)** 'Standard accounts of linguistic competence do not tell the whole story about how human communication succeeds' and **(ii)** 'Standard accounts tell little if any of the truth about how human communication succeeds'. Davidson's thesis is true only if it stops at **(i)**, and is important only if it stretches as far as **(ii)**.)

[1] 'A Nice Derangement of Epitaphs', in R. E. Grandy and R. Warner (eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends* (Oxford University Press, 1985), p. 162.

Desiderata for a theory of meaning

Davidson's theory of meaning is offered as showing how we can discover and express all the facts about the meanings in a language, while respecting the following truths.

(1) The study of meaning must start with whole sentences. Word-meanings are theoretical items that help us to manage and account for the observable facts about what sentences mean.

(2) A normal language user has meanings for indefinitely many sentences, and he doesn't assign them arbitrarily. So he must command a theory that is in a broad, loose sense 'recursive'—a limited set of rules that generate infinitely many consequences of the form 'S means that p'.

(3) What a sentence means is, basically, what somebody means by it.

(4) It is possible to discover empirically what someone means by a given sentence.

(5) There is no sharp line around what a person means by sentence S. If he accepts S' because he accepts S, there may be no fact of the matter about whether for him the truth of S and the falsity of S' is ruled out by sheer meanings or whether instead his beliefs about the world come into it. Davidson attributes this thesis that 'behavioural or dispositional facts. . . on which a theory of interpretation can be based will necessarily be a vector of meaning and belief' to Quine, and says that it is part of 'one of the few real breakthroughs in the study of language' (pp. 148 f.; see also pp. 27, 62, 197).

(6) There is no worthwhile answer to the question 'What is meaning?' or 'What is it for S, as used by speaker x, to mean that p?' This might seem to subvert the whole endeavour, but really it doesn't, as will appear.

Outline of Davidson's theory

The mainspring of Davidson's theory is the notion of a T-sentence, that is, a sentence of the form

S is true (in the idiolect of speaker x) if and only if p.

He holds that there are empirical ways of sorting T-sentences out into true and false, whereas it is not clear how to sift the true from the false among sentences of the form

S means (in the idiolect of speaker x) that p.

He also thinks that T-sentences are clean and decent in a way that sentences using the unvarnished 'mean' are not. But the facts about what sentences mean can be made empirically accessible and conceptually manageable, Davidson says, if they are approached through T-sentences.

One link between the two is obvious. Anyone who is willing to speak of 'meaning' at all, and who thinks that sentences have truth values, will agree that if the meaning of S is truly reported in something of this form:

S (as used by x) means that p,

then the corresponding T-sentence,

S (as used by x) is true if and only if p,

must also be true. The converse doesn't hold, however, for the T-sentence 'Gravity obeys an inverse square law' is true if and only if the speed of light is finite is true because both its clauses are true, but the corresponding meaning sentence

'Gravity obeys an inverse square law' means that the speed of light is finite

is patently false. But Davidson thinks that the members of a certain privileged subset of T-sentences do generate corresponding truths about meanings, and are indeed the source of the whole truth about the meanings of the sentences in a language. 'What I call a theory of meaning has after all turned out to make no use of meanings. . . [but it] supplies all we have asked so far of a theory of meaning' (p. 24).

Because he will not use the concept of meaning, regarding it as unfit for serious use except when based on his T-sentence approach, Davidson does not argue that every T-sentence in the subclass corresponds to a truth about meanings. Rather, he sees this as a conjecture, to be tested against our untutored intuitions about meaning. Of the privileged T-sentences he says that 'we can say that' by giving one of them 'we give [the] meaning' of its subject sentence (p. 56 n.), and of the words 'are true if and only if' in a T-sentence he says that 'we may interpret them if we please as meaning "means that".' (p. 60).

Getting at meaning: demonstratives

As I have remarked, Davidson aims to get facts about what sentences mean—or 'interpretations', as he often says—out of only a subset of T-sentences:

> A theory of truth will yield interpretations only if its T-sentences state truth conditions in terms that may be treated as 'giving the meaning' of object language sentences. Our problem is to find constraints on a theory strong enough to guarantee that it can be used for interpretation. (p. 150)

In the upshot, Davidson has three ways of narrowing down the field. (In recently added footnotes on p. 26, all three are emphasized, but conjunctively, with no suggestion that they are connected as I think two of them are.)

Many of the sentences whose meanings are in question are demonstrative, that is, they contain indexicals like 'here' and 'now'. The T-sentences for these cannot generate truths about meanings in quite the advertised way. For example, the T-sentence

> "It is raining here now' as used by x at t is true if and only if it is raining where x is at t

is all right, but it must not lead us to conclude things like:

> 'It is raining here now' as used by Charles at noon means that it is raining where Charles is at noon.

That is not right, because Charles may not know that he is Charles, or that it is noon when he speaks. As Davidson says, in T-sentences about demonstrative sentences (or 'demonstratives', for short) 'the right side of the biconditional never translates the sentence for which it is giving the truth-conditions' (p. 175; see also pp. 35, 74f.).

Still, T-sentences about demonstratives can still be informative about their meanings, and the right side of any of them will be 'systematically related' (p. 46) to the meaning of the sentence named on the left side. Indeed T-sentences about demonstratives cannot be as remote from the meanings of their topic sentences as can ones about non-demonstratives. For a typical English speaker x we can truthfully complete this:

> 'It rains somewhere at some time' (as used by x) is true if and only if. . .

in madly irrelevant ways, such as

> . . . there is more than one galaxy
> . . . diamonds are harder than glass

and so on. But if we start off with

> 'It is raining here now' (as used by x at t) is true if and only if. . . where x is at t,

we cannot easily make this true without putting into the blank something that means the same as 'it is raining'. As Davidson says: 'A theory that makes the right sentences true at the right times for the right speakers will be much closer to a theory that interprets the sentences correctly than one that can ignore the extra parameters' (pp. 74f).

That seems right: we get at the meaning of a demonstrative sentence by finding out what is in common to all the person-time pairs for which it is true and to none for which it isn't. But that does not apply to sentences containing no

indexicals. For example, we cannot get at the meaning of

It did, does, or will rain somewhere at some time

by examining the spatiotemporal zones at which it is true, because it is true everywhere and always.

Still, our findings about the meanings of demonstratives could lead us to conclusions about the meanings of the other sentences in the language. Using the circumstantial facts about some demonstratives to assign meanings to them as semantic lumps, we could then generate some theory—what Quine calls 'analytical hypotheses'—about how the meanings of those sentences result from the meanings of their parts and the significance of how they are combined. Then we could go on to assign meanings to non-demonstrative sentences built out of some of the same parts.

Sometimes Davidson himself seems to entertain such a picture:

> The first step... settles matters of logical form. The second step concentrates on sentences with indexicals... The last step deals with the remaining sentences... (p. 136; see also p. 168)

but in this work as a whole indexicals seem to me not to get the primacy they deserve. I guess that this is because Davidson got off on the wrong foot, starting with the idea of a T-sentence as something of the form

[Sentence name] is true if and only if [sentence],

where the sentence on the right *translates* the one named on the left. When the named sentence is a demonstrative, that doesn't work, as Davidson himself points out; but he does not scrap his original starting point and start afresh. On the contrary, he stays faithful to the 'translation' version of T-sentences,[1] and so he cannot put demonstratives at the centre of the stage. He characterizes them as a 'very large fly in the ointment' (p. 33), as 'a tricky matter' (p. 131), as involving a 'radical conceptual change' in the program he started out with (p. 58), and speaks of 'adjusting' his 'theory of truth' to accommodate them (p. 213). In 'In Defence of Convention T' (paper 5) they are described as an 'important, indeed essential, factor in making a truth theory a credible theory of interpretation' (p. 74), so that things cannot be 'as I have been pretending' (p. 75); but this is when they are introduced, for the first time in that paper, on its last page. They are given strong primacy in 'True to the Facts', but not in the context of Davidson's theory of meaning (pp. 43 f.).

Getting at meaning: holism

Davidson offers two other ways of delimiting T-sentences so that each of them generates a truth about the meaning of the subject sentence, whether or not it is demonstrative. They correspond to two ways of understanding the requirement that the T-sentence's truth not be 'accidental'.

One thing he means by that (e.g. on p. 175) is that the T-sentence must not merely be true but must be generated by a systematic, comprehensive, 'reasonably simple' (p. 26n.) theory about the totality of sentences in the idiolect under study. Suppose that the idiolect is normal English, and that we have somehow established interpretative T-sentences for some of its sentences, so that S_1's meaning can be recovered from

S_1 is true if and only if p_1

and S_2's from

S_2 is true if and only if p_2

and so on. Now consider the great range of sentences of the form S_j-'and'-S_k. We can systematically generate T-sentences for all of these, so long as we already have them for the separate clauses, through the one formula or

[1] Perhaps because it is the peg on which he hangs the name of Tarski—an irrelevance which I shall discuss later.

T-sentence schema:

> S_j-'and'-S_k is true if and only if (p_j and p_k).

That illustrates the sort of thing Davidson means by T-sentences that are generated by theory, and why he thinks that if T-sentences are arrived at like that they will be apt to strike us as interpretative, i.e. as saying what their subject sentences mean. Thus:

> We can interpret a particular sentence provided we know a correct theory of truth that deals with the language of the sentence. For then we know not only the T-sentence for the sentence to be interpreted, but we also 'know' the T-sentences for all the other sentences; and of course, all the proofs. Then we would see the place of each sentence in the language as a whole, we would know the role of each significant part of the sentence, and we would know about the logical connection between this sentence and others. (pp. 138f.; see also pp. 61, 73f)

Davidson also contends that we can home in on correct interpretations not only by moving from one sentence to a whole idiolect, but also in moving from an individual's idiolect to the largest dialect of which it is a typical part (pp. 152f).

Both contentions are highly plausible, and I think they are true. It would be good to have them defended and explained, however, if only to help us to understand why we find them plausible. On p. 74 Davidson asks why, but what follows is no answer, so far as I can see; and elsewhere he doesn't even raise the question. I think that a proper defence of the holism constraint—that is, of the thesis that a good theory implying T-sentences for every sentence in a language will thereby imply facts about what the sentences mean—would have to draw on Davidson's third way of trying to get a theory of T-sentences to state the facts about meanings. Let us now look at that.

Getting at meaning: counterfactuals

When Davidson speaks of a T-sentence as not 'accidentally' true, he sometimes means not that it flows from a theory but rather that counterfactuals flow from it:

> Sentences of the theory are empirical generalizations about speakers, and so must be not only true but lawlike. "'Snow is white" is true if and only if grass is green' presumably is not a law, since it does not support appropriate counterfactuals. (p. 26n., added in 1982)

At that point he does not say what sorts of counterfactuals, but an answer can perhaps be figured out from this:

> Given that the evidence for this law, if it is one, depends ultimately on certain causal relations between speakers and the world, one can say that it is no accident that 'Schnee ist weiss' is true if and only if snow is white; it is the whiteness of snow that *makes* 'Schnee ist weiss' true. (p. xiv)

This rests on the idea that the truth of a sentence in x's idiolect ultimately rests on x's relating to it in a certain way. From that it follows that some counterfactual of the form

> If snow were not white, x would not have relation R to 'Schnee ist weiss'

entails the counterfactual

> If snow were not white, 'Schnee ist weiss' (as used in x's idiolect) would not be true,

or, for that matter,

> It is because snow is white that 'Schnee ist weiss' (as used in x's idiolect) is true.

(When Davidson writes that what makes the sentence true is 'the whiteness of snow', I don't take him to mean that the sentence is made true by a property. He must mean that it is

made true by *snow's being white*, i.e. by *the fact that snow is white*. (Not that he would put it like that; see pp. 70, 194.)

How do we discover that S is true?

If we know what constraints a system of T-sentences for x's idiolect must obey if it is to generate all the truths about what x's sentences mean, the next thing we need is some way of empirically testing systems of such sentences. This approach can lead us to a defensible theory of meaning for x's idiolect only if we can tell whether a given T-sentence is true or not, without yet knowing what its subject sentence S means; so we must often be able to tell whether S (as used by x) is true, without knowing what it (as used by x) means. One might think that 1 can only learn that

> 'Gravity obeys an inverse square law' (as used by you) is true

by learning what you mean by that sentence and then discovering whether what you mean is true, that is, whether that proposition is true. If that were right, then Davidson's program would be doomed.

Davidson, however, thinks we can get evidence that S (as used by x) is true, in advance of knowing what it means, by two steps: 'The fact that speakers of a language hold a sentence to be true (under observed circumstances) [is] prima facie evidence that it is true under those circumstances' (p. 152); and 'We can know that a speaker holds a sentence to be true without knowing what he means by it or what belief it expresses for him' (p. 162). Let us look at these in turn.

True and held true

Davidson insists that as an interpreter one must assume a large measure of agreement between one's own beliefs and those of the person whose idiolect one is studying. He assumes (I think rightly) that in interpreting the mind of another creature one ought in general to attribute what one takes to be error only if one can explain why it occurs; such explanations cannot be solid unless one already has some grounded theory about the subject's mind; and so in starting to establish such a theory one would be well advised to keep error out of the story at the outset, and to see where one can get by assuming that the subject's beliefs are all true. That seems to be the spirit of this:

> We want a theory that. . . maximizes agreement, in the sense of making [the subject] right, as far as we can tell, as often as possible. . . Once the theory begins to take shape it makes sense to accept intelligible error and to make allowance for the relative likelihood of various kinds of mistake. (p. 136)

In fact, the error-ignoring device is merely the simplest of many possible ways of getting the theory started. But Davidson doesn't see it like that, because he thinks of it as more than a convenient device. His method, he says, involves

> assigning truth conditions to alien sentences that make native speakers right when plausibly possible, according, of course, to our view of what is right. What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. (p. 137)

The claim that in the absence of massive agreement the notion of (dis)agreement is not 'intelligible' goes far beyond merely advising theory builders to start with a working assumption of massive agreement. Davidson repeats this often:

> We will have to assume that in simple or obvious cases most of his assents are to true, and his dissents from false, sentences—an inevitable assumption since the alternative is unintelligible. (p. 62)

This requires not merely that the interpreter of x share most of x's beliefs, but that he share most of x's beliefs about the truth-values of sentences. I find even the former claim, let alone the latter, quite implausible. What is unintelligible in the idea that some creature's perceptual bad luck and intellectual frailty should coincide so as to make most of its beliefs about 'simple or obvious' matters false? It would be different if Davidson argued strongly for his view, but he doesn't. The best he offers is this:

> Widespread agreement is the only possible background against which disputes and mistakes can be interpreted. Making sense of the utterances and behaviour of others, even their most aberrant behaviour, requires us to find a great deal of reason and truth in them. To see too much unreason on the part of others is simply to undermine our ability to understand what it is they are so unreasonable about. (p. 153)

The final sentence of that, vague as it is, has some chance of being right, but only because it sticks with 'reason' (see also pp. 159 f.) and drops 'truth'.

Incidentally, all of that concerns **i** the thesis that x cannot be understood by y unless most of x's beliefs are y's too, that is, are *judged true by y*. At one point Davidson moves from that to the thesis **ii** that x cannot be understood (by anyone) unless most of his beliefs are *true*. It seems prima facie that **i** can be true and **ii** false: a mostly false corpus of beliefs might be understood, on the basis of complete agreement, by an interpreter whose own beliefs were mostly false. But Davidson argues, through a bold transcendental argument, that **i** implies **ii**:

> There is nothing absurd in the idea of an omniscient interpreter; he attributes beliefs to others, and interprets their speech on the basis of his own beliefs, just as the rest of us do. Since he does this as the rest of us do, he perforce finds as much agreement as is needed to make sense of his attributions and interpretations; and in this case, of course, what is agreed is by hypothesis true. But now it is plain why massive error about the world is simply unintelligible, for to suppose it intelligible is to suppose there could be an interpreter (the omniscient one) who correctly interpreted someone else as being massively mistaken, and this we have shown to be impossible. (p. 201)

This is employed in 'The Method of Truth in Metaphysics' (paper 14) to throw a firm, broad bridge between linguistic premises and metaphysical conclusions. It is an amazingly ambitious endeavour. In fact, though, Davidson has not 'shown' that **i** is true; he has merely said that it is. And his perfectly valid transcendental argument, run contrapositively, tells strongly against **i**. The thought of an omniscient interpreter reminds us of the implausibility of the claim that *any* interpreter of x's thought and speech must share most of x's beliefs.

Davidson has another direct argument for the thesis **ii** that any understandable belief system is mostly true:

> We can. . . take it as given that most beliefs are correct. The reason for this is that a belief is identified by its location in a pattern of beliefs; it is this pattern that determines the subject matter of the belief, what the belief is about. Before some object in, or aspect of, the world can become part of the subject matter of a belief (true or false) there must be endless true beliefs about the subject matter. False beliefs tend to undermine the identification of the subject matter. (p. 168)

I don't follow this unless it is relying on a Fregean view about how a thought grips onto a particular, assuming that the only way a creature could believe something about a particular is by having a belief of the form: The thing that is G and H and

I and J and K. . . is also F.[1] If it is relying on that, that is a reason to reject it.

Of course, Davidson's programme for meaning theory doesn't require him to hold that any interpreted belief system is mostly true,[2] or even that y can interpret x only if he shares most of x's beliefs. All he needs is the safer thesis that if we can learn enough about what sentences x holds to be true (in what circumstances), we can move from that to well-grounded conclusions about which of his sentences actually are true. I think that the safer thesis is true, though my reasons for it are ones that Davidson could not accept. How he could defend it, I do not know.

However, even if we hand that to Davidson on a plate, we have still left dangling the question of how we are to discover facts of the form 'x holds S to be true at t'. Let us now turn to that.

Holding true

In one place where Davidson writes that 'we can sometimes tell that a person accedes to a sentence we do not understand', he adds that the line he is taking 'obviously owes its inspiration to Quine's account of radical translation in Chapter II of *Word and Object*' (p. 27). Even in 1967 when he wrote that, Davidson was probably not buying into Quine's account of how we can discover whether a speaker holds S to be true, namely: putting S to him interrogatively, and seeing whether he behaves affirmatively. A dozen years later, anyway, he explicitly disavows this aspect of Quine's account: 'Where [Quine] likes assent and dissent because they suggest a behaviouristic test, I despair of behaviourism and accept frankly intensional attitudes towards sentences, such as holding true' (p. 213).

What does Davidson put in place of Quine's behaviouristic Yes/No way of discovering what sentences the subject assents to? A possible position is this: 'Quine is basically right, but his supposedly decisive binary test for assent to sentences is too simple. In practice, assent shows up in behaviour that is more various, more complex, harder to describe, than Quine allows for. It would be a waste of time to try to describe them in detail.' There is evidence that Davidson would agree with that, but his departure from Quine goes deeper: unlike Quine, he thinks that the route from the behavioural basis through to conclusions about meaning must run through 'frankly intensional attitudes such as assent', which he equates with the notion of x's believing a sentence to be true.

There is a prima facie difficulty for Davidson here. He has insisted that precise and finely shaved information about x's beliefs and intentions cannot be gathered from his behaviour unless this includes interpreted speech, and has claimed this to be fatal to Grice's attempt to explain meaning in terms of belief and intention; so he needs to explain why his own use of the concept of belief, in the notion of 'holding true' on which his account of language is founded, is not killed by the same bullet. Here is part of his answer to that:

> A good place to begin is with the attitude of holding a sentence true, of accepting it as true. This is, of course, a belief, but it is a single attitude applicable to all sentences, and so does not ask us to be able to

[1] In 'Reality Without Reference' (paper 15) Davidson responds to critics who complain that his theory of meaning does not include a (correct, non-Fregean) theory of reference. Davidson's response to this complaint, starting at p. 222, ought to throw light on the passage I have just quoted, but after many readings I still do not understand it.

[2] But that thesis is dear to him, it seems. He returns to its defence in 'A Coherence Theory of Truth and Knowledge', in *Kant oder Hegel?*, Klett-Cotta, 1983.

> make finely discriminated distinctions amongst beliefs. (p. 135)

The claim is that Davidson's prelinguistic use of the concept of belief is trouble-free, when others are not, because the former does not involve any troublesome 'finely discriminated distinctions amongst beliefs'.

Why doesn't it? Davidson's answer to this (pp. 163 f.) is sketchy, to put it mildly, but I think the following is a fair reconstruction of it.

If we are tempted to say that the dog thinks there is a cat up the tree, we encounter questions that we cannot answer unless the dog can speak. **i** How does it think of the cat—as a furry enemy? as an animal of about its own size? as the object it chased yesterday? **ii** What does it believe about the cat? That it is up a tree? that it is out of sight above the ground? that it is where it was when chased yesterday? If we cannot answer questions like these, we do not know what belief it would be right to attribute to the dog. In contrast with this, **i** we can credit x with believing something about the sentence S, without having to face nasty questions about how it thinks of S—as the F sound, as the G sound, as the sound that x just made. And **ii** we can credit x with believing of S that it is true, without being challenged by rival accounts according to which x believes S to have some other property than truth though coinciding with it in this instance. Let us examine these two claims in turn.

i Since any sound has innumerable properties, don't we have the problem of determining which of them enter into x's belief about it? Yes, but if the sound is a sentence token the problem can be solved for it as it cannot for the cat. Every sentence token falls under a strong, general theory that attends to a determinate subset of its features (the ones that will appear in the dictionary and the grammar book, if we ever get x's idiolect interpreted); and we can discover which features those are, and will then be entitled to suppose that if x thinks about the noises at all he does so under that subset of their features. This part of what I conjecture to be Davidson's position seems to be all right.

ii But I can find no defence for the other part. I see no reason to think that, at a stage where we don't yet understand any of x's sentences, we are free from troubles about 'finely discriminated distinctions' between believing S to be true, believing it to be plausible, believing it to be probable, believing it to be desirable, and so on. Indeed, it looks as though there will be troubles about coarsely discriminated distinctions between beliefs about truth and beliefs about properties that have nothing to do with truth. To be entitled to suppose that what x believes about S is that it is true, we must have evidence that x's belief fits S into x's value system in a certain manner: it is of the essence of the concept of truth that truth is what we go in for.[1] After the one sentence that I have quoted, Davidson devotes to this vitally important matter one sentence more. I shall examine it in due course.

Summary up to here

Davidson has argued that when you are interpreting the mind and speech of another creature x, you are entitled to assume that most sentences that x accepts as true are true in his idiolect; and that you can discover what sentences x accepts without knowing what they mean. If he is right on both counts, you can find out which of x's sentences are true and which are not (either absolutely or at specified place-times). Then you need to inform yourself about the rest of the world, discovering whether it is the case (either

[1] Thus Michael Dummett, 'Truth', *Proceedings of the Aristotelian Society* 54 (1958–9).

absolutely or at specified place-times) that p, for various values of p. After that, all you need is a tiny amount of logical skill—just enough to be able, when you know whether S is true in x's idiolect and you know whether p, to know whether

> S is true in x's idiolect if and only if p.

These modest materials, then, will let you sort out true from false amongst all such sentences, and amongst their more complex cousins where the subject sentence contains indexicals.

The true T-sentences thus established will include all kinds of true junk, such as

> 'Snow is usually black' is true in x's idiolect if and only if there are three even prime numbers,

as well as ones that embody interpretations of the subject sentences. But we can filter out the junk. If we establish a set of T-sentences in which **i** many concern demonstrative sentences, **ii** each is implied by a strong, comprehensive theory about all the sentences in x's idiolect, and (iii) each remains true when 'if and only if' is replaced by 'if (and because), and only if', then the T-sentences in that set will correspond to truths about what x's sentences mean. Or so Davidson says, and I agree.

My exposition should make clear why Davidson's approach to meaning, if it works at all, satisfies the desiderata **(1)**, **(3)**, **(4)** and **(6)** with which I started. I may need to say a little about desideratum **(5)**, that is, about how Davidsonian meaning theory relates to Quine's thesis that what a speaker conveys through the sheer meaning of what he says shades into what he conveys through firmly held associated beliefs. Suppose we know that

> Because snow is white, Kurt holds 'Schnee ist weiss' to be true,

or the corresponding counterfactual

> If snow were not white, Kurt would not hold 'Schnee ist weiss' to be true.

That bears on what he means by that sentence, but it does not pretend to decide between

> **(a)** He means by it that snow is white

and

> **(b)** He means by it that snow is F for some F distinct from whiteness but believed by Kurt to be causally necessary and sufficient for whiteness.

Quine's thesis that pure meaning cannot be extracted from the meaning-belief complex is, precisely, the thesis that no decision should be made between **(a)** and **(b)**, the apparent difference between them being an illusion produced by bad philosophy; and Davidson's approach to meaning beautifully respects that view of the matter. In fact, it does so in several ways, but one is enough for now.

Logical form

There remains desideratum **(2)**: a good theory of meaning must be constructive—it must use recursive rules to get unlimited results out of limited basic resources.

Obviously Davidson must envisage his sort of meaning inquiry as being like that: we cannot learn an unlimited language unless it is recursive; and we cannot theorize comprehensively about such a language unless our theory is recursive. The latter claim is true for any sort of theory, and is especially obvious for Davidson's. We could not possibly investigate x's relations with each of the sentences in his idiolect, one by one.

Thus, for example, if we have truth-conditions (that is, T -sentences of the right kind) for n non-conjunctions, then a single rule lets us generate truth-conditions for the 2^n simple conjunctions that can be formed out of them, the 2^{2^n} conjunctions that can be formed out of them, and so

on. Other rules will handle disjunctions, negations, tenses, modals, and so on; and yet others will dig down to the subsentential level, letting us get truth-conditions for large classes of simple sentences out of brute facts about their smaller number of components (pp. 47 f.).

Our theory can have these luxuries only if they correspond to regular features—elements of structure—in the language under study. As Davidson puts it, his sort of 'theory of truth for a language' serves to 'give the meanings of all independently meaningful expressions on the basis of an analysis of their structure' (p. 55).

It is, these days, a humdrum point that any successful meaning theory for a language must find some kind of systematic structure in it. But Davidson seems to hold something further, namely that his kind of theory—with its special emphasis on truth—is apt for revealing structure of one special kind, namely the kind embodied in first-order extensional logic, the logic of truth-functions and quantifiers. I shall discuss this.

There is some reason to hope that most significant structure in any language is of this kind. Davidson plausibly says that the structure or 'logical form' that helps us to understand new sentences is also what helps us to know what entailments hold amongst sentences. If that is right, and if the best meaning theory attributes to the language a structure of truth-functions and quantifiers, then we get the good news that the entailments that hold amongst our informal sentences can be captured in and explained by a system of logic that is powerful, simple, and well understood:

> By abstracting quantificational structure from what had seemed a jungle of pronouns, quantifiers, connectives, and articles, Frege showed how an astonishingly powerful fragment of natural language could be semantically tamed. Indeed, it may still turn out that this fragment will prove, with ingenuity, to be the whole. (p. 51)

So far, so good. But a question remains. Does Davidson think that the adoption of his particular kind of meaning theory somehow guarantees (or at least increases the chances) that we shall find that first-order logic permeates the language under study? His answers to this are given in an uncertain voice.

(Many of them are worded in ways that reflect Davidson's habit of calling his theory of meaning a 'theory of truth' (e.g. on pp. 56, 83, 132, 150, 203). This and some allied phrases are a bar to understanding, and we must get them out of the way if we want a clear, true picture of what happens in Davidson's work on language. Davidson says he is describing a 'recursive account of truth' (p. 57), trying to 'shed light on truth', pursuing a 'systematic account of truth' (p. 62), 'characteriz[ing] truth' (p. 133). He also gives his approach to meaning the label 'Convention T', this being Tarski's name for a certain proposal about what should count as a satisfactory theory of truth. The mentions of Tarski, and the implications that the topic of investigation is truth, have a mainly incantatory role throughout the book, except in 'True to the Facts' **(3)** and 'In Defence of Convention T' **(5)**, neither of which has much to do with the theory of meaning. Davidson implies as much when he writes that 'The interest of a theory of truth, viewed as an empirical theory of a natural language, is not that it tells us what truth is in general, but that it reveals how the truth of every sentence of a particular L depends on its structure and constituents' (p. 218; see also pp. 134, 150). Such remarks, while they remind us that the concept of truth is hard at work in Davidson's theory of meaning, show that the latter is not a theory of truth, is not concerned with defining the truth predicate, and has nothing to do with Convention T. Now back to the main thread.)

Sometimes Davidson seems to hold that his program for investigating meanings merely demands that structure be found and exploited, with no bias towards any particular kind of structure:

> The suggested conditions of adequacy for a theory of truth do not (obviously, anyway) entail that even the true sentences of the object language have the form of some standard logical system. (p. 58)

Also:

> Convention T, in the skeletal form I have given it, makes no mention of extensionality, truth functionality, or first-order logic. . . Restrictions on ontology, ideology, or inferential power find favour, from the present point of view, only if they result from adopting Convention T (p. 68),

that is, only if we are led to them by bringing the T -sentence treatment of meaning to bear on the empirically given language.

Still, he has a view about what we are likely to be driven to if we study meanings in his way. He says that his theoretical constraints 'apparently cannot be met without assigning something very much like a standard quantificational form to the sentences of the language' (p. 132), and that 'If the metalanguage is taken to contain ordinary quantification theory, it is difficult, if not impossible, to discover anything other than standard quantificational structures in the object language' (pp. 150 f.; see also p. 176).

When he writes that 'The semantic constraint in my method forces quantificational structure on the language to be interpreted' (p. 136 n.), I think Davidson overstates his own considered view, which is the more cautious one about what 'apparently' cannot be done.

How do we find structure?

Let us grant that in doing Davidsonian meaning theory we shall be led to attribute to the subject language a structure describable by first-order logic. I now ask: will our Davidsonian activities help us to uncover such structure? When Davidson writes that 'The result of applying the formal constraints. . . is to fit the object language to the procrustean bed of quantificational theory' (p. 151), he suggests that he has been offering a help, a technique for laying bare in the object language those logical forms that correspond to quantificational logic.

But that seems not to be his considered view. Here he divides the whole process into two stages:

> In the first stage, [T-sentences will be given], not for the whole language, but for a carefully gerrymandered part of the language. This part, though no doubt clumsy grammatically, will contain an infinity of sentences which exhaust the expressive power of the whole language. The second part will match each of the remaining sentences to one or. . . more than one of the sentences for which truth has been characterized. (p. 133; see also pp. 29, 203)

It is in the second stage that Davidson would have us matching 'She dried herself with a towel' with 'Something was a drying, was by her, was of her, and was done with a towel', and matching 'Hobbes said that he had a mind to go home' with 'I have a mind to go home. Hobbes said that.' And in this second stage, when we are arriving at and evaluating the likes of Davidson's theories about adverbial modification and oratio obliqua, we have at most two debts to his basic approach to meaning. **(i)** It motivates us to seek and value first-order logical structure in the object language. **(ii)** It liberates us from the view that meaning equivalences are answerable to facts about the conscious

thoughts of speakers, allowing us to maintain that 'She dried herself with a towel' means something about the existence of a drying just so long as the T-sentence

> 'She dried herself with a towel' (as used by us) is true if and only if something was a drying, was by her, was of her, and was done with a towel

satisfies the constraints that Davidson says are required if a T-sentence is to generate a truth about meaning. We should not expect to get from the Davidsonian theory any guidance in actually arriving at T-sentences such as that one. In digging for logical structure in the object language, it seems (see pp. 210–14), we shan't be helped by the idea of a T-sentence, or by the idea of the empirical study of the assents of a speaker, or by anything else that is specially Davidsonian. Or so I conjecture: Davidson does not discuss the question.

A puzzling claim

Davidson sometimes seems to credit his approach to meaning with impressive powers. Here, for example:

> The striking thing about T-sentences is that whatever machinery must operate to produce them, and whatever ontological wheels must turn, in the end a T-sentence states the truth-conditions of a sentence using resources no richer than, because the same as, those of the sentence itself. Unless the original sentence mentions possible worlds, intensional entities, properties, or propositions, the statement of its truth conditions does not. (p. 132)

Davidson doesn't say what it is for a sentence to 'mention' something. Does 'She dried herself with a towel' mention events? Davidson will have to say that it does. So presumably any sentence will count as mentioning Fs if a good translation or interpretation of it, in a clean canonical notation, explicitly quantifies over Fs. But then the 'striking thing' turns out to say only that a good translation of a sentence will not invoke conceptual resources that the sentence itself does not use, which trivially follows from the platitude that a good translation of a sentence will mean the same as the sentence.

What Davidson intended, I think, was not a platitude but a swipe at those who analyse English sentences with help from possible worlds—Montague, Lewis and others. There is no doubt about that intention when the 'striking thing' appears on p. 56, where it is described as a truth 'yet to be made precise'. Davidson goes on to discuss how it should be interpreted—as though its author were someone else—finds 'natural' an interpretation which he says has the effect of 'judg[ing] much recent work in semantics irrelevant to present purposes', and then coolly walks away.

Questions, commands, etc.

Davidson's broad approach to language and meaning has encouraged him to look at some specific problems of semantic analysis, and to suggest solutions. Three papers in this collection tackle such problems. I hope to discuss 'Quotation' **(6)** elsewhere, and 'On Saying That' **(7)** is too well known to need to be displayed here. In any case, although Davidson labels this trio of papers as 'Applications' of his approach to meaning, those two are not tightly related to the central theory. The same is not true of 'Moods and Performances' **(8)**, which tackles a problem that urgently needs solution if Davidson's approach to meaning is to cover the ground. Even if the program as so far expounded is perfectly in order, it leads us only to the meanings of sentences that can be true; so it doesn't touch imperative, optative and interrogative sentences.

In tackling this, Davidson distinguishes kinds of sentence marked off by mood (indicative, imperative etc.) from ways of using sentences (asserting, ordering, etc.), saying that the two classifications are inter-related but are not identical. He rejects, for example, the idea that all there is to giving an order (say) is uttering something in the imperative mood:

> Once a feature of language has been given a conventional expression, it can be used to serve many extra-linguistic ends. . . There cannot be a form of speech which, solely by dint of its conventional meaning, can be used only for a given purpose, such as making an assertion or asking a question. (pp. 113 f.; see also pp. 266–70)

The point is clearly right, and Davidson's presentation of it is powerful and illuminating.

Now, according to Davidson how does a grammatical mood relate to the associated illocutionary act? That is, how does he propose to fit moods other than the indicative into his T-sentence theory of meaning? Here is his answer to that, stated for imperatives (p. 120):

> We can give the semantics of the utterance of an imperative sentence by considering [i] the truth conditions of the utterance of an indicative sentence got by transforming the original imperative, and [ii] the truth-conditions of the mood-setter. The mood-setter of an utterance of 'Put on your hat' is true if and only if the utterance of the indicative is imperatival in force.

On this account, truth-conditions fit in neatly: the mood-setter is true if and only if the utterance has imperatival force, and the content of the imperative is given by an associated indicative S for which there will also be a suitable T-sentence. Uttering the imperative is not, of course, tantamount to uttering the conjunction of the mood-setter and the associated indicative (if it were, then every disobeyed order would be false). But we can understand the imperative by understanding the truth-conditions of the mood-setter and the associated indicative.

This may serve for imperatives and optatives, but it won't work for all interrogatives (as Hintikka has already warned Davidson; see p. 115n). It applies to ones that are used to ask whether—yes or no—it is the case that p. But when it comes to 'Which of them turned off the lights?' or 'What is your name?' the theory won't work: there is no associated indicative that will do the job.

That is a relatively minor blemish, however, compared with Davidson's way of simply helping himself to 'imperatival force' and related notions for the other moods. Uttering a sentence with one force or another—imperatival, assertive, questioning, etc.—involves relating to it in one way or another; and Davidson tells us nothing about any of these relations. You might think that this passage speaks of them:

> For the sake of the present discussion at least we may depend on the attitude of accepting as true, directed to sentences, as the crucial notion. (A more full-blooded theory would look to other attitudes towards sentences as well, such as wishing true, wondering whether true, intending to make true, and so on.) (pp. 195 f.)

But it doesn't. Accepting as true a sentence that one utters is not the same as asserting its truth, though the two are inter-related, as we find when we try to explain what assertion is. Wanting a sentence that one utters to be true is not the same as uttering it with imperative force, though again relations between the two emerge when we try to explain what it is for a particular act to be an imperative. And so on through the others. Davidson encounters none of these relations because he does not dig at all into the concepts of assertion, command etc. To do so, he would have change his

ground radically, consenting to develop a generic theory of belief and intention and then present meaning as a species within that.

Davidson's primitives

At one dramatic point the notion of assertion, or something like it, bears the whole weight of the theory. In defending his claim that without knowing what x's sentences mean we can learn which of them he believes to be true—a claim without which the entire project collapses—Davidson says two things, of which I have discussed one. In the other he says that believing-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. (p. 135)

Given the failure of the first defence, everything rests on this second one.

Davidson here proposes to base conclusions about what sentences x believes to be true on premises about when x 'intends to express a truth in uttering a sentence', adopting the complex notion of *intending to express a truth*, or perhaps the related notion of assertion, as an unanalysed, unexamined lump. Thus, his entire case for saying that his project is empirically feasible, which depends on his view that we can learn that x holds S to be true before we have interpreted S, comes to rest on a fragile, complex structure that Davidson has taken out of the box and set in place without subjecting it to the least analytical scrutiny.

This silence about what assertion is, or about the notion of intending to express a truth, is part of a larger silence. Repeatedly in this book Davidson relies on the premise that *he speaks a language*, but he never subjects that to any kind of explanation or analytical scrutiny. Indeed, with one strange exception which I shall discuss shortly, he tells us nothing about what it is for a behavioural system to be a language, or for a sound or movement to be (a token of) a sentence. I mean that as a criticism, because I think it is part of a philosopher's task to take warm, familiar aspects of the human condition and look at them coldly and with the eye of a stranger. Indeed, Davidson himself is good at doing this—but never with the concept of language.

His willingness to take that concept on trust, as something whose instances are dropped into our laps without the need for philosophical work, must be operating in the sentence where Davidson helps himself to the notion of uttering S 'intending to express a truth'. Looked at from a proper analytical distance, this is a slapdash, careless, unthorough performance. But Davidson is not at that distance. He stands in the thick of the human situation, helping himself to things that he finds within reach—things like the concept of language, of sentence, of intention to express a truth.

Starting with belief and desire

If Davidson's meaning-theory were based on, rather than mixed up with, a theory of beliefs and desires, it might be able to complete itself, with a decent account of holding true, and of what it is for a behavioural system to be a language and for a sound or movement to be a sentence.

There are ways of discovering what a creature wants and what it thinks, and some work has been done on laying out criteria by which one moves from **(i)** premises about what someone does (output) in relation to what he experiences (input) to **(ii)** conclusions about what he thinks, and from those to **(iii)** further conclusions about what he means by his utterances. The route from **(ii)** to **(iii)** was found by Grice.

As well as offering some prospect of foundations, a Gricean approach also improves on something lying right at the heart of the limited theory that Davidson does offer. He says that

> 'Snow is white' as used by x is true if and only if snow is white

is informative about meaning only because it is non-accidentally true, which means that

> It is because snow is white that x holds 'Snow is white' to be true.

What sort of causal transaction is this? It has the same form as

> It is because snow is cold that x is putting on his muffler;

and we naturally want in the former case what we have in the latter, namely some idea of what the causal route is from the fact about snow to the fact about x. How does snow's being white lead to x's having that attitude to that sentence?

It seems indisputable that at least part of the story runs as follows. We start with

> **(1)** snow's being white;

this leads causally to

> **(2)** x's thinking that snow is white;

and this combines with x's taking 'Snow is white' to mean that snow is white to cause

> **(3)** x's thinking 'Snow is white' to be true.

Davidson can say all of this. He does say: 'A sentence is held true because of two factors: what the holder takes the sentence to mean, and what he believes' (p. 167), and he would surely say that when someone assents to 'Snow is white' because snow is white, the causal chain runs through his believing that snow is white rather than through the other component, namely his meaning by 'Snow is white' that snow is white.

There is, however, something funny about how Davidson has to tell this story. When he implies that the causal route runs from **(1)** to **(3)** via **(2)**, with extra input from the fact that x means by 'Snow is white' that snow is white, this is just a theory-based gloss on the situation after the **(1)**-**(3)** link has been established. By Davidson's own lights, we could not have discovered the bit of the chain that runs from **(1)** to **(2)**, independently discovered what x means by 'Snow is white', and on this basis predicted that x would hold that sentence to be true. This is impossible, according to Davidson, because in his scheme of things we need the facts about which sentences x holds to be true as our basis for judgments about what x takes various sentences to mean.[1]

In contrast with this, the Gricean approach lets us take the causal chain one step at a time. Rather than having to start with the great leap from snow's being white to x's having a certain attitude to the sentences, and then breaking it down into smaller steps in accordance with a certain theory about what must be going on, the Gricean theory offers us a chance of being able to take the chain one step at a time: we learn about what causes x to acquire what beliefs, we learn about what he means by various sentences, and those two discoveries put us in a position to predict that he will regard the sentence 'Snow is white' as true.

That strikes me as preferable, if it is possible. Perhaps Davidson would agree, but he doesn't think it is possible. Let us see why.

One reason is as follows. The Gricean handling of this matter brings in meaning en route to the T-sentence, by basing it on belief and intention; and Davidson holds that

[1] And for our judgments about what x believes? One might think so from some of what Davidson says, but we shall see that it is not really so.

that must be wrong,

> because of. . . what may be called *the autonomy of meaning*. Once a sentence is understood, an utterance of it may be used to serve almost any extralinguistic purpose. An instrument that could be put to only one use would lack autonomy of meaning; this amounts to saying it should not be counted as a language. . . It is largely this that explains why linguistic meaning cannot be defined or analysed on the basis of extralinguistic intentions and beliefs. (pp. 164f.; see also pp. 274f.)

But he has introduced this as a prima facie difficulty for his own approach, which he then meets quite satisfactorily. He doesn't observe that the Gricean can follow suit. The Gricean needs only one-way conditionals of the form

> If R(x,S) then S as used by x means that P,

where R(x,S) is strong enough to make the conditional true but weak enough to allow it to be interesting and instructive, even if not weak enough to make its converse true as well. This one-way Gricean approach lets us get a theory of meaning launched for a given language, while leaving the door open for x to exploit his meaning for S in ways not captured by the relation R.

Davidson could respond that that is to give up the idea that 'linguistic meaning can be *defined* or *analysed* on the basis of extralinguistic intentions and beliefs'. And so it is, if an analysis must be expressible as an analytic biconditional. But Davidson's own position remains in peril unless the Gricean approach in general—and not just the biconditional version of it—can be rejected. Let us see what else he has to say against it.

As I have already mentioned, Davidson also objects against Grice's program that the needed information about beliefs and intentions cannot be gathered until we know some meanings:

> Radical interpretation cannot hope to take as evidence for the meaning of a sentence an account of the complex and delicately discriminated intentions with which the sentence is typically uttered. . . We cannot hope to attach a sense to the attribution of finely discriminated intentions independently of interpreting speech. The reason is not that we cannot ask the necessary questions, but that interpreting an agent's intentions, his beliefs and his words are parts of a single project, no part of which can be assumed to be complete before the rest is. If this is right, we cannot make the full panoply of intentions and beliefs the evidential basis for a theory of radical interpretation. (p. 127; see also pp. 143f and 163f.)

Griceans are here invited into the trap of allowing that their evidential basis must be 'complete' before any conclusions are drawn from it. Why should they fall into that?

They can keep out of it by taking the line that some inquiry into beliefs and intentions can yield some preliminary hunches about meanings, that these can add starch to further inquiries of the former sort, resulting in more and better conclusions about meanings, and so on upwards by the bootstraps. This version of the Gricean program is modest enough to escape both of Davidson's objections, yet contentful enough to be a real rival to his approach. Furthermore, it respects all six of the desiderata for a theory of meaning listed near the start of this paper.[1]

[1] Grice's own version does not respect the sixth desideratum. Where I offer only one-way conditionals, from belief and intention to meaning, he tries for analytic biconditionals.

Thought without language

Grice's approach takes meaning as a species of intending. This lets us **i** see our linguistic behaviour as one kind of intentional behaviour, and **ii** see creatures with language as one kind of creatures that behave intentionally. Davidson would not regard **ii** as an advantage, because he confines intentionality to the tiny fragment of the animal kingdom that understands some language. This comes up repeatedly, especially in 'Thought and Talk' (paper 11), whose 'chief thesis' is 'that a creature cannot have thoughts unless it is an interpreter of the speech of another' (p. 157).

What reasons does Davidson give for this bold hypothesis? In 'Thought and Talk' two are mentioned with some sympathy.

(i) If a creature cannot speak, there is a limit to how determinately fine-grained our attributions of mental content to it can be. This is true, though there is less to it than Davidson seems to think. Having exhibited some of the difficulties, he says that they are even worse when it comes to universal thoughts, conditional thoughts, 'or thoughts with, so to speak, mixed quantification ("He hopes that everyone is loved by someone")' (p. 164). That last example may owe more to 'hopes' and to 'loves' than to mixed quantification as such. Change the example to 'He thinks that everyone has got something to eat', and the search for behavioural evidence looks less hopeless.

Davidson seems unsure how much force to give his point about 'fine distinctions'. On p. 164 he says that it 'do[es] not constitute an argument' for denying thought entirely to speechless creatures, and he is surely right about that. Later on he writes that because 'fine distinctions between beliefs are impossible without understood speech', 'we must have a theory that simultaneously accounts for attitudes and interprets speech, and which assumes neither' (p. 195), but that is patently wrong: we could instead have a theory that starts by coarsely interpreting attitudes, and from there makes the first small steps towards interpreting speech. I prefer to hold Davidson to his wiser, earlier stand.

(ii) On the last page of the paper, Davidson embarks on his one convinced argument for its 'chief thesis'. It starts with the premise that 'Belief. . . as a private attitude. . . is not intelligible except as an adjustment to the public norm provided by language' (p. 170). Davidson is not saying that the concept of belief can be justifiably applied only on the strength of linguistic evidence (and if he did say that, it would put him in trouble with some other things he says). His thesis is the weaker one that the concept of belief can rightly be applied only within an inquiry or program of study in which it is sometimes applied on the evidence of linguistic behaviour. On a reasonable understanding of this premise, it is still strong enough to imply the 'chief thesis' that a creature cannot have thoughts unless it belongs to a speech community.

What are Davidson's grounds for the premise? Astonishingly, he offers none. In the preceding paragraphs, he has been expounding his view that we can interpret x's speech only on the assumption that x is not guilty of massive error, and has been showing how occasional attributions of error

[1] [The key to this footnote is on the next page.] Davidson's phrase about 'the *public* norm provided by language' might mean only that language is outer, or it might be meant to confine the discussion to dialects shared by many people. The latter reading is encouraged by this: 'Belief is built to take up the slack between sentences held true by individuals and sentences true (or false) by public standards' (p. 153). But I don't think Davidson consideredly holds that belief is built only to do that. He says later: 'The concept of belief. . . stands ready to take up the slack between objective truth and the held true' (p. 170), which has nothing to do with relating an individual to his society.

can still help to protect a good interpretative theory for x's language from being too easily refuted.[1] The story goes like this: if the theory implies that

> S as used by x is true if and only if p,

and on some occasion we find that x holds S to be true although in fact not-p, we can preserve the theory by supposing that this time x is in error, which frees us in this instance to break the link between 'x holds S to be true' and 'S as used by x is true'.

This is a perfectly good story, as far as it goes. The concept of belief does involve the concept of error, and the latter is serviceable in the interpretation of speech. But Davidson goes on from that to assert without argument that the concept of belief comes 'only' from this way of using it, and that belief is 'not intelligible except as an adjustment to the public norm provided by language'. This performance is the book's low point, on the scales of cogency and of courtesy; I am at a loss to account for it.

Although Davidson does not argue for his premise, I shall argue against it. Here is a sketch of another context in which the concept of belief is intelligible:

> We formulate hypotheses about x's goals, wanting to use them to explain some of his movements. If we tried to do this in terms of the concept of what *would in fact* realize his goals, we would find that we couldn't get a theory that was at once decently comprehensive and nearly true. So we introduce the concept of belief, allowing for both error and ignorance. We hypothesize that he does what *he believes will* realize his goals, and we save this from vacuity by developing some theory about what sorts of error or ignorance he is likely to have in what sorts of situation.

This is a prima facie intelligible use of the concept of belief, and is not confined to the study of creatures that speak.[2] Davidson has provided me with no reason to suppose that I am mistaken about this and that the concept of belief can be intelligibly used only in contexts where speech is being interpreted.

Incidentally, it is because Davidson ties belief so tightly to language that he can say that 'Error is what gives belief its point' (p. 168). Really it gives it half of its point, the other half coming from ignorance: as well as 'x believes that p, though really not-p' we have 'x does not believe that p, though really p'. The two can be put on a par by a theory that grounds belief in the explanation of non-linguistic behaviour, but not by one that ties belief to language in the way Davidson's does. We need the concept of x's ignorance to explain why x does not act on the fact that there is something edible behind that stone, but we need not appeal to x's ignorance to explain why x does not say that there is something edible behind that stone. The plausibility of the idea that error and ignorance are twins is a further reason for rejecting Davidson's approach to belief.[3] But my main point is just that Davidson's tying of belief to language is not obviously right and is not supported by any argument at all.

Belief and the concept of belief

Once he has arrived at that position, Davidson's use of it as a premise is peculiar. The natural development is this:

[2] In this criticism of Davidson's position, I don't focus on Davidson's tying of the concept of belief to the idea that x interprets the speech of others. Davidson insists on this, as against the idea that x speaks—'What is essential to my argument is the idea of an interpreter, someone who understands the utterances of another'—but I cannot see why. Perhaps he is running it together with his weaker claim that 'a creature must be a member of a speech community if it is to have the concept of belief'. Even that latter claim is too strong to represent Davidson's best opinion.

[3] For more on error and ignorance, see my *Linguistic Behaviour*, section 14, and *A Study of Spinoza' s Ethics*, section 40.

> **(1)** Belief is intelligible only as an adjustment to the public norm provided by language. It follows that **(2)** a creature must be a member of a speech community if the concept of belief is to be intelligibly applied to it.

It seems clear that **(1)** does entail **(2)**, which is the declared chief thesis of the paper. But what Davidson says is this (the numbers are mine):

> **(1)** Belief. . . is not intelligible except as an adjustment to the public norm provided by language. It follows that **(2*)** a creature must be a member of a speech community if it is to have the concept of belief.

It's odd that instead of stepping forward from **(1)** to **(2)**, which is his ultimate conclusion that the creature must have language if it is to have beliefs, Davidson steps sideways from **(1)** to (2*), which says that the creature must have language if it is to have the concept of belief, thus creating a need for a further step from that to **(2)**. I am at a loss to explain this detour.

The step from **(1)** to (2*) seems all right, but the further step to **(2)** is not. Here it is:

> Can a creature have a belief if it does not have the concept of belief? It seems to me that it cannot, and for this reason. [i] Someone cannot have a belief unless he understands the possibility of being mistaken, and [ii] this requires grasping the contrast between truth and error—true belief and false belief. But [iii] this contrast, I have argued, can emerge only in the context of interpretation, which alone forces us to the idea of an objective, public truth. (p. 170)

Anyone who is persuaded by this must be taking **[i]** at first weakly enough to be acceptable, and then more strongly so as to imply **[ii]**. If there is any truth in **i**, it is just this: a creature cannot think that p unless it can think that perhaps not-p, and in that sense understand the possibility of being mistaken. If x thinks the cat is up the tree, he must understand the possibility that the cat is not up the tree. But in **[ii]** Davidson assumes that if x thinks the cat is up the tree he must grasp the possibility that he is wrong in thinking that the cat is up the tree. I cannot imagine anyone's accepting this unless he had been deceived by an ambiguity in **[i]**.

So presumably the argument collapses, because **[ii]** is supposed to conjoin with benum[iii])—though I don't understand how—to imply that a creature cannot have a belief without having the concept of belief.[1]

So Davidson has not made much of a case for holding that 'the notion of a true belief depends on the notion of a true utterance',[2] and he has never addressed himself to any of the literature that purports to launch the former notion without help from the latter.

It matters whether that literature is basically sound, because if we cannot establish concepts of belief and desire in advance of any meaning theory, the whole project of cognitive ethology is doomed. Another result would be that Grice's theory of meaning, which analyses 'By doing A, x means that p' as meaning that x does A intending in a certain complex

[1] Davidson may have been seduced in another way into narrowing the range of the concept of belief. He holds not only (a) that the concept of belief can have no life of its own independently of the concept of meaning, but also (b) that there is no determinate line between what one means by a sentence and what lies outside the meaning while being tied to it by firmly held biconditionals. These two are utterly different: (b) does not speak about whether *belief* is usable apart from *meaning*, but merely says how the two concepts are inter-related in contexts where they are both being used. But perhaps Davidson tends to think of (b) as supporting (a), for example when he advises us to 'think of meanings and beliefs as interrelated constructs of a single theory' (p. 147; see also p. 196), which could proceed from either (a) or (b).

[2] Still less for the remarkable claim that 'Speaking a language is not a trait a man can lose while retaining the power of thought' (p. 185).

way to get someone to believe that p, is condemned to a kind of epistemic circularity or ungroundedness.

But suppose that languageless creatures can believe and intend, so that there is some hope for cognitive ethology; it is still a further question whether such a creature could be known to have beliefs and intentions of the complex sort required for Grice's analysis. If Davidson had to give up his opposition to languageless belief as such, he might still deny that the Gricean program can be empirically grounded. The volume under review retains his 1974 assertion that

> We sense well enough the absurdity in trying to learn without asking him whether someone. . . intends, by making certain noises, to get someone to stop smoking by that person's recognition that the noises were made with that intention. (p. 144)

This is offered, without supporting argument, as 'a principled, and not merely a practical, obstacle' to the explanation of meaning in terms of Gricean intentions. Well, in my *Linguistic Behaviour* (1976) I have described evidence, and a train of reasoning, that could entitle us to attribute a Gricean intention to a creature whom we did not know to have language. Or so I say. There is, at least, a case to be answered.

Incurably foreign languages

Paper 13, 'On the Very Idea of a Conceptual Scheme', throws a bright light on the matters I have been discussing. In it, Davidson attacks the idea that there could be what I'll call an 'incurably foreign language'—one structured so differently from ours that no significant range of its sentences could be translated into ours. His argument for the impossibility of such a language is illuminating.

In the absence of translatability, Davidson asks, what is to qualify an item as a language? For a believer in incurably foreign languages,

> The idea. . . is that something is a language, and associated with a conceptual scheme, whether we can translate it or not, if it stands in a certain relation (predicting, organizing, facing, or fitting) to experience (nature, reality, sensory promptings). The problem is to say what the relation is, and to be clearer about the entities related. (p. 191)

Davidson has damaging things to say about some of these, but his argument weakens when he comes to the idea of a language as something that fits the world, that is, as a system of sentences many of which are true (p. 194)· We cannot 'divorce the notion of truth from that of translation', he says, because Tarski's Convention T, according to which

> a satisfactory theory of truth for a language L must entail, for every sentence S of L, a theorem of the form 'S is true if and only if p' where 'S' is replaced by a description of S and 'p' by S itself if L is in English, and by a translation of S into English if L is not in English,

'embodies our best intuition as to how the concept of truth is used' and thus commits us to tying truth to translation.

On the most natural reading of it, that version of Convention T implies that a satisfactory theory of truth for L must be stated with reference to some particular language (English in the given example, but it could be some other). How could any particular language be conceptually involved? If my concept of truth involves translatability into English, and Boris's involves translatability into Russian, doesn't that imply that we have different concepts of truth? No, for Davidson could hold that each person's concept of truth brings in a particular language or a particular small set of languages—because *each person's concept of truth is partly self-referential*. The idea is that we all have the same concept

of truth, and that it involves the concept of *a language I know*. And that is indeed Davidson's view of the matter: he says that Convention T, in suggesting 'an important feature common to all the specialized concepts of truth', makes 'essential use of the notion of translation into a language we know' (pp. 194 f.).

We must presume that Davidson doesn't just believe this but finds it intuitively plausible. How else can we explain his offering it with no argument beyond the assurance that it 'embodies our best intuition as to how the concept of truth is used'? Speaking for myself, I find it implausible. I am not sure that my concept of truth involves translatability at all; but if it does, it's only translatability into some language. I have no trouble conceiving of creatures whose language cannot be translated into any language that I could understand.

Even if Davidson were right in holding that the concept of truth involves the notion of a language I know, his present use of this as a premise is peculiar: he is explaining *language* in terms of *true*, and explaining *true* in terms of *language I know*. That points up the fact that while in this one context he is circumspect about the concept of language in application to strangers, he has no inhibitions or cautions with respect to the notion of *my language*: he takes this as a conceptual given, an atom rather than a molecule with the general concept language as a component. The streak of incurious parochialism that runs through Davidson's work is nowhere more vividly present than in his way of arguing that nothing could be a language if we couldn't translate it.

* * * *

Re-reading and thinking about these papers has not dislodged my ancient opinion that Davidson's approach to meaning is inferior to that of Grice's papers and the books by Schiffer and myself. I announce that prejudgment loudly, as a warning that I may have been blinkered by it. I would add that if some of my criticisms show that I have missed the significance of certain passages, I don't take all the blame for that. This book does not easily yield its content to the reader.[1]

[1] I am grateful for good help from William P. Alston, Simon Blackburn, Thomas McKay, Shekar Pradhan, Alexander Rosenberg, and Robert Van Gulick.