7

Theories of truth

1 A summary sketch[1]

The object of this section is to sketch the main kinds of theories of truth which have been proposed, and to indicate how they relate to each other. (Subsequent sections will discuss some theories in detail.)

Coherence theories take truth to consist in relations of coherence among a set of beliefs. Coherence theories were proposed e.g. by Bradley 1914, and also by some positivist opponents of idealism, such as Neurath 1932; more recently, Rescher 1973 and Dauer 1974 have defended this kind of approach. *Correspondence* theories take the truth of a proposition to consist, not in its relations to other propositions, but in its relation to the world, its correspondence to the facts. Theories of this kind were held by both Russell 1918 and Wittgenstein 1922, during the period of their adherence to logical atomism; Austin defended a version of the correspondence theory in 1950. The *pragmatist* theory, developed in the works of Peirce (see e.g. 1877), Dewey (see e.g. 1901) and James (see e.g. 1909) has affinities with both coherence and correspondence theories, allowing that the truth of a belief derives from its correspondence with reality, but stressing also that it is manifested by the beliefs' survival of test by experience, its coherence with other beliefs; the account of truth proposed in Dummett 1959 has, in turn, quite strong affinities with the pragmatist view.

[1] Proponents of the theories I shall discuss take different views about what kinds of items are truth-bearers. In what follows I shall speak variously – depending upon which theory I am discussing – of 'beliefs', 'sentences', 'propositions' etc. as true or false; only when the difference makes a difference shall I draw attention to it.

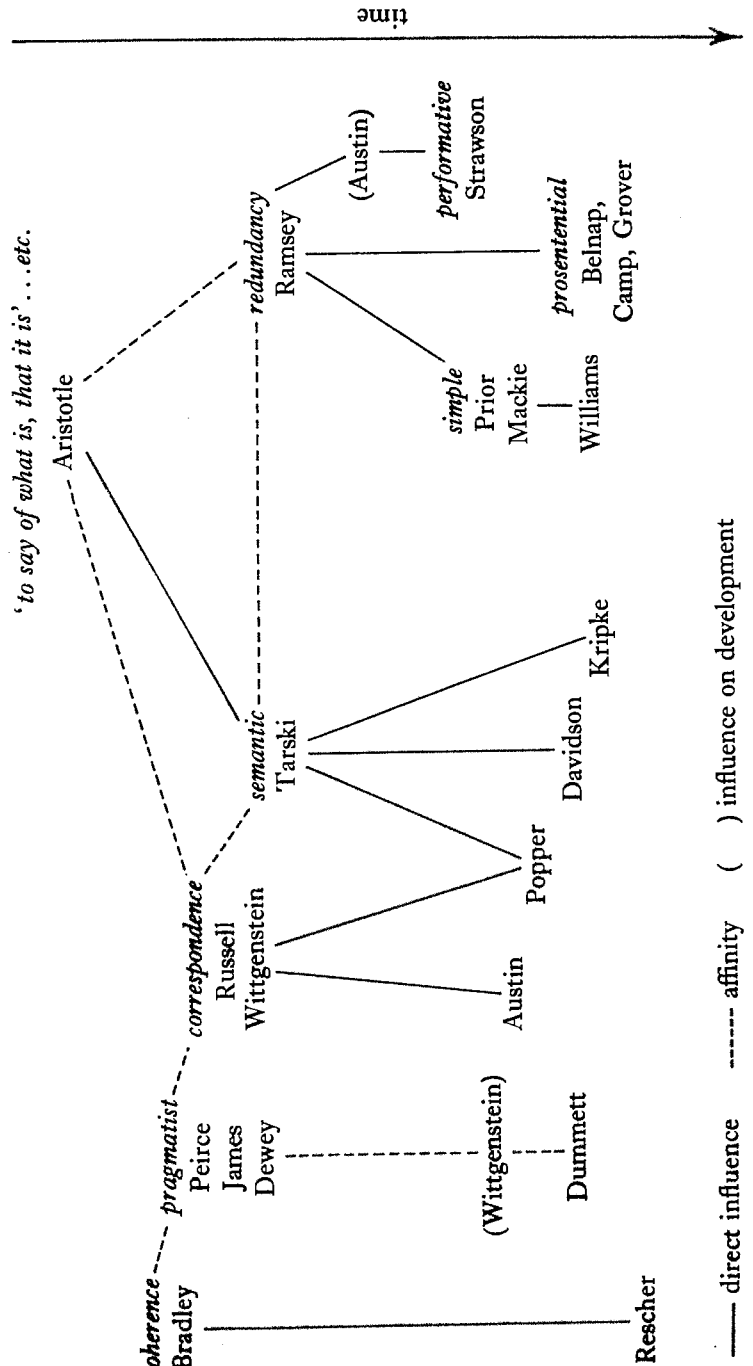


Fig. 4 Theories of truth

Aristotle had observed that 'to say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true'. In proposing his *semantic* theory of truth, Tarski 1931, 1944 aims to explicate the sense of 'true' which this dictum captures. Truth, in Tarski's account, is defined in terms of the semantic relation of satisfaction, a relation between open sentences (like '$x > y$') and non-linguistic objects (such as the numbers 6 and 5). The truth theory recently proposed by Kripke 1975 is a variant of Tarski's, modified essentially to cope in a more sophisticated way with the semantic paradoxes. Popper's account of truth and his theory of verisimilitude or nearness to the truth is based upon Tarski's theory, which Popper regards as supplying a more precise version of traditional correspondence theories.

The *redundancy* theory of truth, offered by Ramsey 1927, claims that 'true' is redundant, for to say that it is true that p is equivalent to saying that p. It is evident that this account has some affinities with Aristotle's dictum, and consequently with some aspects of Tarski's theory. There have been several recent variants of Ramsey's theory: Strawson's 'performative' account (1949); the 'simple' theory of truth suggested by Prior 1971 and amplified by Mackie 1973 and Williams 1976; and the 'prosentential' theory presented by Grover, Camp and Belnap 1975.

Definitions versus criteria of truth

A distinction is commonly made (by e.g. Russell 1908b, Rescher 1973 ch. 2, Mackie 1973 ch. 1) between *definitions* of truth and *criteria* of truth; the idea is, roughly, that whereas a definition gives the meaning of the word 'true', a criterion gives a test by means of which to tell whether a sentence (or whatever) is true or false – as, for example, one might distinguish, on the one hand, fixing the meaning of 'feverish' as having a temperature higher than some given point and, on the other, specifying procedures for deciding whether someone *is* feverish.

This distinction needs careful handling. One's suspicions may be aroused by the existence of disagreement about which theories of truth count as definitional and which as criterial: for instance, while Tarski himself disclaims any interest in supplying a criterion of truth, and Popper regards it as an advantage of the semantic theory that it is definitional rather than criterial, Mackie counts Tarski's theory as – and criticises it for – aspiring to supply a criterion. And one's suspicions should be confirmed by some clearly inappropriate uses of the distinction. For example, Russell accused the pragmatists of having confused the definition and the criterion of truth, when the pragmatists held that the meaning of a term is correctly given precisely by supplying criteria for its application. (It is not at all unusual, I'm afraid, for a philosopher who deliberately identifies *A*s and *B*s to find himself facing the criticism that he has 'confused' *A*s and *B*s.)

However, one cannot simply decide to refrain from using the distinction, problematic as it is, because of its importance to such questions as whether the coherence and correspondence theories need be regarded as rivals, between which one is obliged to choose, or as supplementing each other, correspondence supplying the definition and coherence the criterion. This question is at issue even between proponents of the coherence theory. Thus Bradley, conceding that 'Truth to be truth must be true of something, and this something itself is not truth' (1914 p. 325), seems to allow that an account of the meaning of truth may require appeal to something like correspondence, while coherence is rather a mark, a test, of truth. Blanshard, by contrast, insists that truth *consists in* coherence, which is a definition as well as a criterion. This insistence seems to be based on the conviction that there must be some intimate connection between a dependable criterion and what it is a criterion *of*. Coherence could not be the test, but correspondence the meaning, of truth, he argues, for then there is no explanation why coherent beliefs should be the ones that correspond to the facts; if coherence is to be a reliable test of truth, it must be because it is constitutive of the meaning of truth (see Blanshard 1939 p. 268).

Rescher proposes (1973 chs. 1 and 2) to deflect this argument by distinguishing between *guaranteeing* (infallible) and *authorising* (fallible) criteria, and arguing that only in the case of guaranteeing criteria need there be the connection with the definition that Blanshard thinks inevitable. This distinction illuminates some issues touched upon earlier. Rescher counts C as a guaranteeing criterion of x if:

necessarily (C iff x obtains)

But – as Rescher observes – in this sense, any definition of truth would also supply an infallible criterion of truth. For instance, if truth consists in correspondence to the facts, then, necessarily, if 'p' corresponds to the facts, 'p' is true, so correspondence is an infallible

criterion.[1] (The idea that Tarski gives a criterion of truth may derive from this conception of criteria.)

So: if one has a definition, one thereby has a 'guaranteeing' criterion. The converse, however, is a bit less straightforward. It is a guaranteeing criterion of a number's being divisible by three, for example, that the sum of its digits be divisible by three, but this, I take it, is not what it means for a number to be divisible by three. Rather: if one has a guaranteeing criterion, then either it *is*, or else it is *a logical consequence of*, a definition.

An authorising criterion, however, is fallible: it is not necessarily the case that (C iff x obtains); so, either it is true, though not necessarily, that C iff x obtains, or perhaps it is not invariably true that C iff x obtains. (Rescher considers the second but not the first kind of case.) So an authorising criterion of x is distinct from a definition of x – it need not be logically related to the meaning of 'x'.[2]

But now why, if any definition supplies a guaranteeing criterion, should one ever want an authorising criterion? The answer is rather clear, I think, but hard to put precisely: if one wants to find out whether x obtains, one would like, ideally, a reliable indicator of the presence of x which is *easier to discover to obtain* than x itself. A definition gives an indicator which is perfectly reliable, but exactly as difficult to discover to obtain as x itself; an authorising criterion gives an indicator which may be less than completely reliable, but which, by way of compensation, is easier to discover to obtain. For example, one might think of the characteristic spots as an authorising criterion of measles; not a foolproof test, since it is not logically necessary that one has the spots iff one has the measles, but much more easily discovered than, say, the presence of a given bacterium which is (or

[1] If one identifies meaning and criterion – as the pragmatists do – then one is obliged to hold the criterion to be guaranteeing. This will be pertinent to the discussion in §6 below of Popper's argument that the pragmatist theory of truth threatens fallibilism.

[2] Rescher does not explicitly amplify the 'necessarily' in his account of a guaranteeing criterion, but contextual clues indicate that he has in mind logical necessity, which is the interpretation I have used. If physically necessary tests were included, the previous and some subsequent paragraphs would have to be rewritten to allow criteria which are related to that of which they are a test by physical necessity, as well as criteria related by logical necessity, to count as guaranteeing. Of course, the distinction between logical and physical necessity – and indeed, the distinction between the necessary and the contingent – is not unproblematic.

so I shall suppose for the sake of argument) the guaranteeing criterion.

So far, then, Rescher's vindication of Bradley's view of coherence as a criterion (an authorising criterion, that is) but not a definition, of truth, against Blanshard's argument for an inevitable connection between definition and criterion, is successful. However, it is pertinent that a weaker version of Blanshard's idea seems to work even for authorising criteria. It seems plausible to argue that, if C is an authorising criterion (even in the least favourable case when its presence isn't invariably correlated with that of x), then there ought to be *some* kind of connection – not, indeed, a logical connection, but perhaps a causal connection, for example – between x and C. Consider, again, spots as an authorising criterion of measles; there is a causal connection between the spots and the disease of which they are the symptom. And, indeed, this is relevant to a feature of Bradley's account which Rescher neglects. It is plausible to think that Bradley believed there to be a connection between one's beliefs' being coherent, and their corresponding to reality (i.e. between the authorising criterion and the definition), for he holds that reality is coherent.

The concept of truth is as important to epistemology as to philosophy of logic. Some theories of truth have an important epistemological component, are concerned with the accessibility of truth; and the search for a criterion of truth is often a manifestation of this concern. It is noticeable that on the whole the theories on the left side of the sketch of truth-theories (fig. 4) take the epistemological dimension more seriously than those on the right, with the coherence and pragmatist theories epistemologically rich, but redundancy theories, at the other extreme, with virtually no epistemological 'meat' (as Mackie puts it) on them.

2 Correspondence theories

Both Russell and Wittgenstein, during their 'logical atomist' periods,[1] offered definitions of truth as the correspondence of a proposition to a fact.

Propositions, according to Wittgenstein, are verbal complexes; molecular propositions (such as '$Fa \vee Gb$') are composed truth-functionally out of atomic propositions (as, 'Fa'). The world consists

[1] Wittgenstein was the originator of logical atomism, but Russell's version appeared first, in his 1918 lectures, while Wittgenstein's was presented in 1922 in the *Tractatus*.

of simples, or logical atoms, in various complexes, or arrangements, which are facts. And in a perfectly perspicuous language the arrangement of words in a true, atomic proposition would mirror the arrangement of simples in the world; 'correspondence' consists in this structural isomorphism. The truth-conditions of molecular propositions can then be given; '$-p$' will be true just in case 'p' is not true, '$p \vee q$' will be true just in case either 'p' is true or 'q' is true, and so forth.

Wittgenstein's version of logical atomism is austere; Russell augmented it with an epistemological theory according to which the logical simples about whose character Wittgenstein is agnostic are sense-data, which Russell took to be the objects of direct acquaintance, and the meaningfulness of a proposition is supposed to derive from its being composed of names of objects of acquaintance. These epistemological additions do not vitally affect the core of the account of truth; but some other differences between Russell's and Wittgenstein's versions are more relevant. Russell's account has the virtue of recognising the difficulties in regarding all molecular propositions, notably belief propositions and universal quantifications, as truth-functions of atomic propositions. Other features of Russell's version, however, seem to create unnecessary difficulties; for instance, he allows (though with less than complete confidence, because of the adverse reaction this thesis received at Harvard!) negative as well as positive facts, so that the truth of the negation of p can consist in its correspondence to the fact that not p, rather than p's failure to correspond to the facts; and the suggestion that there are two correspondence relations, one of which relates true propositions and the other false propositions to the facts, seems gratuitous, indeed, in view of the admission of negative facts, doubly so.

Numerous critics have observed that the trouble with the correspondence theory is that its key idea, correspondence, is just not made adequately clear. Even in the most favourable cases the required isomorphism between the structure of a proposition and that of the fact involves difficulties; consider:

The cat is to the left of the man (the proposition)

(the corresponding fact)

even here (as Russell concedes, pp. 315–16) it looks as if the fact has

two components, the proposition at least three; and of course the difficulties would be much severer in other cases (consider 'a is red', 'a is married to b', or for that matter 'the cat is to the right of the man'). The interpretation of correspondence as a structural isomorphism is intimately connected with both the theory about the ultimate structure of the world and the ideal of a perfectly perspicuous language, characteristic theses of logical atomism. The question arises, therefore, whether the correspondence theory can be divorced from logical atomism, and, if it can, what account could then be given of the correspondence relation.

Austin 1950 supplies a new version of the correspondence theory, a study of which offers some answers. Austin's version does not rely either on an atomist metaphysics or on an ideal language; the correspondence relation is explicated, not in terms of a structural isomorphism between proposition and fact, but in terms of purely conventional relations between the words and the world. Correspondence is explained via two kinds of 'correlation':

(i) 'descriptive conventions' correlating words with *types* of situation

and

(ii) 'demonstrative conventions' correlating words with *specific* situations

The idea is that in the case of a statement such as 'I am hurrying', uttered by s at t, the descriptive conventions correlate the words with situations in which someone is hurrying, and the demonstrative conventions correlate the words with the state of s at t, and that the statement is true if the specific situation correlated with the words by (ii) is of the type correlated with the words by (i). Austin stresses the conventional character of the correlations; *any* words could be correlated with *any* situation; the correlation in no way depends on isomorphism between words and world.

A difficulty with this account of correspondence, which essentially appeals to *both* kinds of correlation, is that it applies directly only to statements made by sentences which are indexical, since the demonstrative conventions would have no role to play in the case of sentences like 'Julius Caesar was bald' or 'All mules are sterile', which can't be used in statements referring to different situations. (Austin's comments on these cases, p. 23n, are none too reassuring.)

On the other hand, Austin's version, I think, makes an improve-

ment on Russell's account of 'the facts'. The point is hard to put clearly, but it is significant enough to be worth putting even somewhat vaguely. Russell is apt to speak as if the truth of *p* consists *in its correspondence to the fact that p*; but the trouble with this is that the relation between '*p*' and the fact that *p* is just *too* close, that '*p*' couldn't fail to correspond to *that* fact. His evasiveness about the criteria of individuation of facts may indicate that he felt this discomfort. Austin's version, however, locates the truth of the statement that *p* not in its correspondence to the fact that *p*, but rather in *the facts*' being as '*p*' says, or, as Austin puts it, in the demonstrative conventions' correlating '*p*' with a situation which is of the kind with which the descriptive conventions correlate it. (Austin is aware of this difference; see 1950 p. 23; and cf. Davidson 1973 and O'Connor 1975.)

3 Coherence theories

A coherence theory of truth was held by the idealists (I shall discuss Bradley's account, but related views were held by his German philosophical ancestors Hegel and Lotze) and also by some of their logical positivist opponents. So the relation between coherence theories and idealism is rather like that between correspondence theories and logical atomism – in that in each case the theory of truth became divorced from the metaphysical outlook with which it was originally characteristically associated.

It will be useful – because this way some significant relations between coherence and correspondence theories can be highlighted – to begin in the middle, with Neurath's defence of a coherence view. A little history will not go amiss: the logical positivists, under the influence of Wittgenstein's *Tractatus*, originally subscribed to a correspondence view of the character of truth. They were, however, strongly motivated by epistemological concerns, and consequently desired a test (authorising criterion) of truth – a way to tell whether or not a sentence indeed corresponds to the facts. Carnap and Schlick tackled the problem in two parts; statements reporting immediate perceptual experience, they argued, are incorrigible, that is to say, we can directly verify that they correspond to the facts, and the truth of other statements can then be tested by means of their logical relations to these. Already a characteristic feature of the correspondence theory – that truth lies in a relation between beliefs and the world – is modified: the test of the truth of all but perceptual statements derives from their relations with other statements, the perceptual ones, which are supposed to be verified by direct confrontation with the facts. Neurath, however, raised doubts about the supposed incorrigibility of 'protocols', and having thus denied the possibility of a direct check of even perceptual beliefs' correspondence to the facts, held the only test of truth to consist of relations among beliefs themselves. Our search for knowledge requires a constant readjustment of beliefs, the aim of which is as comprehensive a belief set as consistency allows. (This is strongly reminiscent of the 'method of maxima and minima' in James' epistemology (James 1907); Quine's position in 'Two dogmas of empiricism' (1951), where he endorses Neurath's metaphor of the process of acquisition of knowledge as repairing a raft while afloat on it, is similar. Cf. Hempel 1935 for an excellent account of the development of the positivists' view of truth, and Scheffler 1967 ch. 5 for a lively 'blow by blow' report of the controversy between Schlick and Neurath.)

Neurath's final position has much in common with Bradley's account of the test of truth as 'system', which he explains as requiring both *consistency* and *comprehensiveness* of the belief set. And in Bradley as in Neurath the appeal to coherence is connected with the denial that our knowledge has any incorrigible basis in the judgments of perception. However, Bradley's theory has intimate connections with his absolute idealism. Briefly and roughly, reality, according to Bradley, is itself essentially a unified, coherent whole. (Russell's pluralistic logical atomist metaphysic was motivated by reaction against the idealists' monism.) And while Bradley conceded something to the idea of truth as correspondence to reality, he held that, strictly speaking, nothing short of the fully comprehensive, consistent belief set at which we aim is really true; at best, we achieve partial truth – *part* of the truth is not fully true. The point of these remarks is to bring out a point anticipated earlier (§1) – that the connections between Bradley's view of truth and his view of reality are close enough for it to be somewhat misleading simply to regard him as offering coherence as the test, while leaving correspondence as the definition, of truth; rather, the explanation of the success of coherence as the test derives from an account of reality as itself essentially coherent.

A persistent difficulty with the correspondence theory, as I observed above (§2) has been the difficulty of supplying a precise account of 'corresponds'. A similar problem dogs the coherence theory; it needs

to be specified exactly what the appropriate relations between beliefs must be for them to be 'coherent' in the required sense. Unsympathetic critics of coherence theories – Russell, for example – have been apt to assume that simple consistency is sufficient; Bradley, however, was already insisting (as early as 1909, against criticism from Stout; see Bradley 1914) that comprehensiveness as well as consistency is required.

Rescher, who defends a coherentist epistemology (coherence as the test of truth) offers a detailed explication of the twin requirements of 'system'; consistency and comprehensiveness. The problem facing the coherentist, as Rescher sees it, is to supply a procedure for selecting, from incoherent and possibly inconsistent data ('truth-candidates', not necessarily truths) a privileged set, the warranted beliefs, those one is warranted in holding true. A 'maximal consistent subset' (M.C.S.) of a set of beliefs is defined thus: S' is an M.C.S. of S if it is a non-empty subset of S which is consistent, and to which no member of S not already a member of S' can be added without generating an inconsistency. But the data-set is likely to have more than one M.C.S.; this is the basis of Russell's criticism that coherence cannot distinguish the truth from a consistent fairy tale. To avoid this difficulty Rescher proposes that the M.C.S.s of the data-set be 'filtered' by means of a plausibility index, dividing data into those which are, and those which are not, initially plausible, and thus reducing the number of eligible M.C.S.s. However, this may be insufficient to single out a unique M.C.S.; so Rescher recommends the adoption of the disjunction of those M.C.S.s permitted by the plausibility filter.

Though Rescher's work has contributed significantly to the detailed working-out of a coherentist epistemology, difficulties remain. An obvious problem is the specification and justification of the standards of plausibility (Schlick's appeal to the alleged incorrigibility of protocols could be seen as an alternative response to a related difficulty). A less obvious, but also important, difficulty is that the recommended procedure is, so to speak, static in character: it tells one how to select a privileged, 'warranted', subset from an initial set of data, but correspondingly underestimates the importance of seeking *new* data. (Bradley's insistence that only the most fully comprehensive belief set – the whole truth – is strictly speaking true could be seen as a response to this difficulty.) Coherence will surely form part, but not the whole, of a satisfactory epistemology.

Thus far, I have followed Rescher (with some qualifications in Bradley's case) in taking coherence to be intended as a test of truth, as playing an epistemological role, while allowing correspondence the metaphysical part. (Cf. the large role played by coherence in Quine's epistemology, from 1951 to 1970, with his adoption of the semantic definition of truth, 1970 ch. 3). The pragmatists, however, challenge this distinction with their characteristic criterial theory of meaning.

4 Pragmatic theories[1]

Peirce, James and Dewey offer characteristically 'pragmatic' accounts of truth, which combine coherence and correspondence elements.

According to the 'pragmatic maxim' the meaning of a concept is to be given by reference to the 'practical' or 'experimental' consequences of its application[2] – 'there can *be* no difference' as James put it (1907 p. 45) 'that *makes* no difference'. So the pragmatists' approach to truth was to ask what difference it makes whether a belief is true.

According to Peirce, truth is the end of inquiry, that opinion on which those who use the scientific method will, or perhaps would if they persisted long enough, agree. The significance of this thesis derives from Peirce's theory of inquiry. Very briefly: Peirce takes belief to be a disposition to action, and doubt to be the interruption of such a disposition by recalcitrance on the part of experience; nquiry is prompted by doubt, which is an unpleasant state which one tries to replace by a fixed belief. Peirce argues that some methods of acquiring beliefs – the method of tenacity, the method of authority, the *a priori* method – are inherently unstable, but the scientific method enables one to acquire (eventually) stable beliefs, beliefs which will not be thrown into doubt. For the scientific method, Peirce argues, alone among methods of inquiry, is constrained by a reality which is independent of what anyone believes, and this is why it can lead to consensus. So, since truth is the opinion on which the scientific method will eventually settle, and since the scientific method is constrained by reality, truth is correspondence with reality. It also follows that the truth is satisfactory to believe, in the sense that it is stable, safe from the disturbance of doubt.

[1] This section draws upon Haack 1976c.

[2] Peirce stressed the connection of 'pragmatic' with Kant's use of '*pragmatische*' for the empirically conditioned, James the connection with the Greek '*praxis*', action.

James' major contribution was an elaboration on this idea. The advantage of holding true beliefs, he argued, was that one was thereby guaranteed against recalcitrant experience, whereas false beliefs would eventually be caught out ('Experience . . . has ways of *boiling over* . . .', 1907 p. 145). James' account of the way one adjusts one's beliefs as new experience comes in, maximising the conservation of the old belief set while restoring consistency – strikingly like Quine's 1951 view of epistemology – introduces a coherence element. True beliefs, James comments, are those which are verifiable, i.e. those which are, in the long run, confirmed by experience.

Thus far, I have stressed the continuities between Peirce's and James' views, but there are some differences which should be mentioned. First, while Peirce was a realist, James was inclined towards nominalism (cf. Haack 197+), and therefore embarrassed by the possible-but-not-yet-realised verifications to which the view of truth as verifiability committed him; consequently, although in principle he allows that beliefs are true (false) though no one has yet verified (falsified) them, in practice he is sufficiently persuaded of the pointlessness of dwelling on this that he slips into speaking, inconsistently, as if new truths come into existence when beliefs get verified. (The idea that truth is *made*, that it grows, was taken up by the English pragmatist, F. C. S. Schiller.) Second, James often speaks of the true as being the 'good', or the 'expedient' or the 'useful' belief (e.g. 1907 pp. 59, 145). Unsympathetic critics (e.g. Russell 1908b, Moore 1908) have taken James to be making a crass, not to say morally objectionable, identification of the true with the congenial belief. The comments which provoked this critical fury, when taken in context, can often be read, much more acceptably, as pointing to the superiority of true beliefs as *safe from falsification* (cf. James' own defence, 1909 p. 192 – 'Above all we find *consistency* satisfactory'). But James is also making another claim: that since at any given time the evidence available to us may be insufficient to decide between competing beliefs, our choice may depend upon such grounds as simplicity or elegance (1907 p. 142); a claim which does have connections with his 'will to believe' doctrine.

Dewey adopts Peirce's definition as 'the best definition of truth' (1938 p. 345n). He prefers the expression 'warranted assertibility' to 'truth', and adds the thesis that it is precisely warranted assertibility that characterises those beliefs to which we give the honorific title, knowledge (cf. Ayer 1958). Dummett's view of truth, the direct

inspiration for which derives from the work of the later Wittgenstein and from Intuitionism in the philosophy of mathematics, resembles Dewey's in its stress on assertibility; see Dummett 1959.

The main theses of the pragmatic account can be summarised as follows:

truth is:

- the end of inquiry — Peirce, James, Dewey
- *correspondence* with reality — Peirce, James, Dewey
- satisfactory (stable) belief — Peirce, James, Dewey
- *coherence* with experience – verifiability — James, Dewey
- what entitles belief to be called 'knowledge' — Dewey

5 The semantic theory

Tarski's has been, of late, probably the most influential and most widely accepted theory of truth. His theory falls into two parts: he provides, first, *adequacy conditions*, i.e. conditions which any acceptable definition of truth ought to fulfil; and then he provides a definition of truth (for a specified formal language) which he shows to be, by his own standards, adequate. Both parts of this programme will be examined. The detailed statement of the theory is to be found in Tarski 1931; 1944 is a good introduction.

It isn't hard to see why Tarski's theory should have been so influential. For one thing, his adequacy conditions on definitions of truth promise a kind of filter to discriminate, among the embarrassingly numerous theories of truth, those which meet minimal conditions of acceptability, which therefore have some prospect of success. Furthermore, the method employed in Tarski's definition of truth can be applied to a large class of formal languages. But the very features of Tarski's theory which contribute most to its appeal also, as we shall see, create problems for it: can Tarski's adequacy conditions be given independent motivation? and: have his methods any interesting application to the problem of truth for natural languages?

Adequacy conditions on definitions of truth

The problem which Tarski sets himself is to give a definition of truth which is both *materially adequate* and *formally correct*; the first of these conditions sets limits on the possible content, the second on the possible form, of any acceptable definition.

Material adequacy

Tarski hopes that his definition will 'catch hold of the actual meaning of an old notion' (1944 p. 53). However, the 'old' notion of truth is, Tarski thinks, ambiguous, and even doubtfully coherent. So he restricts his concern to what he calls the 'classical Aristotelian conception of truth', as expressed in Aristotle's dictum:

> To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true.

And he proposes, as material adequacy condition, that *any acceptable definition of truth should have as consequence* all instances of the (T) schema:

(T) *S* is true iff *p*

where '*p*' can be replaced by any sentence of the language for which truth is being defined and '*S*' is to be replaced by a name of the sentence which replaces '*p*'. An instance of (T) would be, e.g.:

'Snow is white' is true iff snow is white

where the sentence on the right-hand side is referred to by its 'quotation-mark name' on the left-hand side.

Tarski emphasises that the (T) schema is *not a definition* of truth – though in spite of his insistence he has been misunderstood on this point. It is a *material adequacy condition*: all instances of it must be entailed by any definition of truth which is to count as 'materially adequate'. The *point* of the (T) schema is that, if it is accepted, it fixes not the intension or meaning but the *extension* of the term 'true'. For suppose one had two definitions of truth, D_1 and D_2, each of which was materially adequate. Then D_1 would entail all instances of:

S is true_1 iff *p*

and D_2 all instances of:

S is true_2 iff *p*

so that D_1 and D_2 are co-extensive. Or, to put essentially the same point in another way, the material adequacy condition would rule out certain definitions of truth, those, that is, which did *not* entail instances of the (T) schema.

But exactly what kinds of definition will the material adequacy condition rule out? In answering this question I shall use a weakened version of the criterion: not that all instances of the (T) schema *be deducible from* any acceptable truth definition (Tarski's version), but that the truth of all instances of the (T) schema *be consistent with* any acceptable truth-definition. The reason for this modification is simply that the weakened adequacy condition is much more readily applicable to non-formal definitions of truth. Now it is to be hoped – and perhaps even expected – that it will allow the sorts of definition which have been seriously proposed, and disallow what one might call 'bizarre' theories. But matters turn out rather oddly. Consider the following definition of truth, which seems to me definitely bizarre: a sentence is true iff it is asserted in the Bible. Now it might be supposed that this definition (I shall call it 'D_B' for short) does not entail all instances of the (T) schema, not, for instance:

> 'Warsaw was bombed in World War II' is true_B iff Warsaw was bombed in World War II.

Now it is indeed the case that someone who did not accept D_B might deny:

> 'Warsaw was bombed in World War II' is asserted in the Bible iff Warsaw was bombed in World War II.

But further reflection makes it clear that a proponent of D_B could perfectly well maintain that his definition *does* entail all instances of (T); he may allow that 'Warsaw was bombed in World War II' is true, but insist that it *is* asserted in the Bible (in an obscure passage in Revelation, perhaps), or if he agrees that 'Warsaw was bombed in World War II' is not asserted in the Bible, he will also, if he is wise, maintain the falsity of the right-hand side of the above instance of the schema. So, rather surprisingly, Tarski's material adequacy condition cannot be relied upon to be especially effective in ruling out bizarre truth-definitions.

The material adequacy condition *does*, however, apparently rule out a certain important class of truth theories, those, that is, according to which some sentences (statements, propositions, wffs or whatever) are neither true nor false. For suppose '*p*' to be neither true nor false; then the left-hand side of:

'*p*' is true iff *p*

will be, presumably, false, while the right-hand side will be neither true nor false. So the whole biconditional will be false, or at any rate untrue. (This argument could, however, be avoided if one were prepared to allow that metalinguistic assertions such as ''*p*' is true' might themselves be neither true nor false.) It is arguable that Tarski's material adequacy condition would rule out at least some versions of the coherence theory; arguably it would *not* rule out a pragmatist theory, since the pragmatist view of meaning would rule meaningless any sentences which are neither verifiable nor falsifiable, so that there could be no meaningful but truth-valueless sentences. It certainly seems rather extraordinary to rule non-bivalent theories of truth out of court.

The idea behind Tarski's material adequacy condition is, presumably, that the truth of the (T) schema is so certain and obvious that it is proper that one should feel confident in rejecting any theory of truth which is inconsistent with it. For myself, I find the initial certainty and obviousness of the (T) schema somewhat modified when it turns out that not only some of the seriously propounded theories of truth, but also some very bizarre theories, are consistent with it, while some other serious theories are inconsistent with it (but see Davidson 1973 for a defence of 'convention T').

Formal correctness

The formal requirement which Tarski lays down concerns the structure of the language in which the definition of truth should be given, the concepts which may be employed in the definition, and the formal rules to which the definition must conform.

It is notorious that semantic concepts, incautiously handled, are apt to give rise to paradoxes (e.g. the Liar – 'This sentence is false'; Grelling's paradox – ''not true of itself' is true of itself iff it is not true of itself', and so forth). Tarski investigates the Liar paradox in some detail, and argues that the antinomy arises from the assumptions:

(i) That the language used contains, in addition to its expressions, (*a*) the means of referring to those expressions and (*b*) such semantic predicates as 'true' and 'false'. Such a language Tarski calls 'semantically closed'.
(ii) That the usual logical laws hold.

Being unwilling to reject assumption (ii), Tarski concludes that

a formally correct definition of truth should be expressed in a language which is not semantically closed.

Specifically, this means that the definition of truth-in-O, where O is the *object language* (the language *for* which truth is being defined), will have to be given in a *metalanguage*, M (the language *in* which truth-in-O is defined). The definition of truth will have to be, Tarski argues, relative to a language, for one and the same sentence may be true in one language, and false, or meaningless, in another. The danger of the semantic paradoxes can be avoided by resort to a metalanguage; the Liar sentence, for instance, will then become the harmless 'This sentence is false-in-O', which is, of course, a sentence of M, and consequently not paradoxical. The object/metalanguage distinction is, of course, a relative one, and a whole hierarchy of languages would be required to define truth at every level. Since all equivalences of the form (T) must, by the material adequacy condition, be implied by the definition of truth, M must contain O or translations of all sentences of O as part, plus the means to refer to expressions of O; for instances of (T) have, on the left-hand side, an expression denoting a sentence of O, and, on the right-hand side, a sentence of O or a translation of a sentence of O. Notice that, in specifying, in the metametalanguage, that the metalanguage, M, should contain either the object language, O, itself, or a translation of each sentence of O, semantic notions are employed (explicitly in the latter case, and implicitly in the former, since M must contain the same expressions of O with the same interpretations as they have in O).

It is also required that the structure of O and M should be 'formally specifiable'. For in order to define 'true-in-O' it will be essential to pick out the wffs of O, since these are the items to which 'true-in-O' applies. (This is one of the reasons which Tarski gives for feeling sceptical about the possibility of defining 'true-in-English' – or 'true' for *any* natural language; the sentences of natural languages are not, he thinks, formally specifiable. Later followers of Tarski, notably Davidson, feel more optimistic on this point. It is one I shall need to investigate more closely.)

Tarski also requires that 'the usual formal rules of definition are observed in the metalanguage' (1944 p. 61). These rules include:

(i) no free variable may occur in the *definiens* which does not also occur in the *definiendum*

ruling out e.g. '$Fx = \mathrm{df}(x+y = 0)$', and

(ii) that no two occurrences of the same variable may occur in the *definiendum*

ruling out e.g. '$Fxx = \text{df } Gx$'. Condition (i) prevents definitions which could lead to contradiction; condition (ii) prevents definitions in which the *definiendum* is ineliminable (cf. Suppes 1957 ch. 8).

Any acceptable definition of truth must, then, according to Tarski, satisfy both the material adequacy and the formal correctness conditions. He gives a definition, and shows that it is, by these standards, acceptable.

Tarski's definition of truth

It might be thought that the (T) schema, though not itself a definition of truth, provides an obvious way of giving such a definition. Tarski himself points out that one could think of each instance of (T) as a *partial* definition of truth, in that each instance specifies the truth-conditions of some one specific sentence; so that a conjunction of *all* instances of the (T) schema, one for each sentence of O, would constitute a complete definition. Tarski, however, argues that it is *not* possible to give such a conjunctive definition, for the number of sentences of a language may be infinite, and in this case it is impossible actually to give all the required instances of the (T) schema.

Neither, Tarski argues, can the (T) schema be turned into a definition of truth by universal quantification. It might be supposed that, using on the left-hand side a quotation mark name of the sentence used on the right-hand side, one could straightforwardly generalise to obtain:

(D) (p) ('p' is true_O iff p)

which would apparently constitute a complete definition, and one, furthermore, guaranteed to be materially adequate, since all instances of (T) are instances of it. But Tarski rejects this suggestion because he believes that the result of quantifying into quotation marks is meaningless. For, according to Tarski (and also according to Quine), the expression obtained by writing quotation marks around an expression is an indivisible unit, analogous to a proper name, so that:

Snow is white

is no more part of:

'Snow is white'

than (to use Quine's example) 'rat' is of 'Socrates'. Tarski concedes

that if it were feasible to regard quotation as a function, then (D) would be no less well-formed than e.g.:

(x) ($x^2 = x.x$)

He thinks, however, that there are overwhelming objections to treating quotation as a function, and, in consequence, that (D) is no more well-formed than e.g.:

(x) (Texas is large)

So Tarski thinks that the (T) schema not only is not, but also cannot be turned into a definition of truth. So he constructs his own definition by a more roundabout route. He takes it as a *desideratum* that no semantic terms should be taken as primitive, so that any semantic notion in terms of which 'true' is defined should itself previously be defined. Since he is to define 'true' using the concept of satisfaction, which is a semantic one, this means that he must first define 'satisfies'.

Informal account

The procedure is as follows:

(a) specify the syntactic structure of the language, O, for which truth is to be defined

(b) specify the syntactic structure of the language, M, in which truth-in-O is to be defined; M is to contain
  (i) either the expressions of O, or translations of the expressions of O
  (ii) syntactical vocabulary, including the names of the primitive symbols of O, a concatenation sign (for forming 'structural descriptions' of compound expressions of O), and variables ranging over the expressions of O
  (iii) the usual logical apparatus

(c) define 'satisfies-in-O', and

(d) define 'true-in-O' in terms of 'satisfies-in-O'

Why does Tarski first define 'satisfies'? Well, first, because he considers it desirable to employ, in his definition of truth, no semantic primitives; for he considers that semantic notions are none of them,

pre-theoretically, sufficiently clear to be safely employed. But why 'satisfies'? This is a suitable notion in terms of which to define 'true' because closed, compound sentences are formed out of *open* sentences, rather than closed, atomic sentences. For example, '$(\exists x)\ Fx \vee Gx$' is formed out of 'Fx' and 'Gx' by the operations of disjunction and existential quantification; and the open sentences 'Fx' and 'Gx' are not true or false, but satisfied, or not, by objects. The definition of satisfaction is *recursive* – that is, definitions are given first for the simplest open sentences, and then the conditions are stated in which compound open sentences are satisfied. (The definition could, however, be turned into an explicit one.) This procedure will provide a definition of truth applicable to all sentences of O.

'*Satisfies*': open sentences are not true or false, they are satisfied, or not, by certain things, pairs of things, triples of things, etc. For instance: 'x is a city' is satisfied by London; 'x is north of y' is satisfied by ⟨London, Exeter⟩; 'x is between y and z' is satisfied by ⟨London, Exeter, Edinburgh⟩...etc. ('⟨..., ...⟩' indicates the *ordered n-tuple* of the n items which appear between the pointed brackets.) The order of the items is obviously important, since ⟨London, Exeter⟩ satisfies 'x is north of y' but ⟨Exeter, London⟩ does not. Satisfaction is a relation between open sentences and ordered n-tuples of objects. To avoid the difficulties arising from the fact that open sentences may have 1, 2, ... or *any* number of free variables, Tarski defines satisfaction as a relation between open sentences and *infinite* sequences, under the convention that '$F(x_1 \ldots x_n)$' is to be satisfied by the sequence $\langle O_1 \ldots O_n, O_{n+1} \ldots \rangle$ just in case it is satisfied by the first n members of the sequence; subsequent members are ignored.

The negation of an open sentence S_1 will be satisfied by just those sequences which do not satisfy S_1; and the conjunction of S_1 and S_2 by just those sequences which satisfy S_1 *and* satisfy S_2. The existential quantification of an open sentence will be satisfied by a sequence of objects just in case there is some other sequence of objects, differing from the first in at most the ith place (where the ith is the variable bound by the quantifier) which satisfies the open sentence resulting from dropping the quantifier. For instance, the sequence ⟨England, London, Edinburgh...⟩ satisfies '$(\exists x)$ (x is a city between y and z)' because e.g. the sequence ⟨York, London, Edinburgh⟩ satisfies 'x is a city between y and z'.

'*True*': Closed sentences are special cases of open sentences, those, namely, with *no* free variables. The first member of a sequence, and

all subsequent members, are irrelevant to whether or not the sequence satisfies a 0-place open sentence, i.e. a *closed* sentence. So Tarski defines a sentence as *true just in case it is satisfied by all sequences*, and as *false just in case it is satisfied by none*. This procedure may be made less mysterious by considering an example. The 2-place open sentence 'x is north of y' is satisfied by e.g. all sequences ⟨Edinburgh London, ...⟩, whatever their third and subsequent members. The 1-place open sentence 'x is a city' is satisfied e.g. by all sequences, ⟨Edinburgh, ...⟩ whatever their second and subsequent members. And the (true) 0-place open sentence '$(\exists x)$ (x is a city)' is satisfied by *all* sequences ⟨..., ..., ...⟩, whatever their first and subsequent members; for there is a sequence, ⟨Edinburgh, ...⟩ for instance, which differs from any arbitrary sequence in at most the first place, and which satisfies 'x is a city'. Any closed sentence will be satisfied by *all* sequences or by *none*, and can't be satisfied by some and not others. Consider a rather austere language: first-order predicate calculus without singular terms. In the simplest case, a closed sentence is formed by existential quantification of a 1-place open sentence. Such an existentially quantified sentence is satisfied by an arbitrary sequence only if there is another sequence, differing from it in the first place at most, which satisfies the 1-place open sentence which results from dropping the initial existential quantifier; and so, if the existential sentence is satisfied by *any* sequence, it will be satisfied by *every* sequence. So a closed existential sentence will be satisfied either by all sequences or by none. The negation of a closed existential sentence, by the negation clause of the satisfaction definition, will be satisfied by a sequence iff the negated sentence is not satisfied by that sequence and so, once again, will be satisfied either by all sequences or by none; and similarly for the conjunction of two closed existential sentences, which will be satisfied by a sequence iff both conjuncts are satisfied by that sequence, and so, also satisfied by all sequences or by none. But why is 'true' defined as 'satisfied by all sequences', and 'false' as 'satisfied by none'? Well, consider again the closed sentence '$(\exists x)$ (x is a city)': let X be an arbitrary sequence of objects. By the clause of the definition of satisfaction which covers existentially quantified sentences, X satisfies this sentence iff there is some sequence Y differing from X in at most the first place which satisfies 'x is a city'; now an object o satisfies 'x is a city' just in case o is a city, so there is such a sequence just in case there is some object which is a city. Thus '$(\exists x)$ (x is a city)' is satisfied by all sequences

just in case some object is a city. (Consult Rogers 1963 for further informal discussion of Tarski's definition.)

Two features of Tarski's definition deserve explicit mention at this point. First, it imposes an *objectual interpretation* of the quantifiers; as the previous example indicates, '$(\exists x)\ Fx$' is true iff some object is F. A substitutional interpretation would avoid the need for the detour via satisfaction, for it would permit truth of quantified sentences to be defined directly in terms of the truth of their substitution instances (cf. ch. 4 §1). Second, in his original paper, Tarski gives an *absolute* rather than a *model-theoretic* definition; 'satisfies', and hence 'true', is defined with respect to sequences of objects in the actual world, not with respect to sequences of objects in a model or 'possible world' (e.g. 'there is a city north of Birmingham' is true, absolutely, but false in a model in which the domain is, say, {London, Exeter, Birmingham, Southampton}; cf. pp. 115, 122 below).[1]

Formal account

Tarski gives his definition of truth for a class calculus (the object language), and uses a formalised metalanguage. I shall give, instead, a definition of truth for a more familiar object language, the first-order predicate calculus, and I shall use English plus the object language (cf. (b)(i), p. 105) as metalanguage. This truth-definition will, however, follow Tarski's in all essentials. (It follows Quine's account in 1970 ch. 3 rather closely.)

Syntax of O

The expressions of O are:

variables: $x_1, x_2, x_3 \ldots$ etc.

predicate letters: $F, G \ldots$ etc. (each taking a given number of arguments)

sentence connectives: $-$, &

quantifier: $(\exists \ldots)$

brackets: (,)

[1] In 1957 Tarski and Vaught give a model-theoretic definition. The significance attached to the difference between absolute and model-theoretic definitions will depend, in part, on one's attitude to possible worlds (see pp. 190 ff. below). Those who think of the actual world as just one possible world among others will think of the absolute definition as simply a special case of a model-theoretic definition. However, since a model-theoretic definition uses semantic primitives (the notion of the interpretation of expressions) in the metalanguage, it does not satisfy all the constraints Tarski used in his 1931 paper; and this seems to some (Davidson for example; see below) to be an important reason to prefer an absolute definition.

In terms of this austere primitive vocabulary, of course, the other truth-functions and the universal quantifier can be defined. I am also assuming that singular terms have been eliminated. The advantage of choosing such a minimal vocabulary is, as will become apparent, that it much reduces the work which has to go into the truth definition.

The atomic sentences of O are those strings of expressions which consist of an n-place predicate followed by n variables.

(i) All atomic sentences are well-formed formulae (wffs)
(ii) If A is a wff, $-A$ is a wff
(iii) If A, B are wffs, $(A \ \& \ B)$ is a wff
(iv) If A is a wff, $(\exists x)\ A$ is a wff
(v) nothing else is a wff

Definition of 'satisfies'

Let X, Y range over sequences of objects, A, B over sentences of O, and let X_i denote the ith thing in any sequence X.

Then satisfaction can be defined for atomic sentences, by giving a clause for each predicate of the language.

1. for 1-place predicates:
   for all i, X: X satisfies 'Fx_i' iff X_i is F
   For 2-place predicates:
   for all i, X: X satisfies 'Gx_ix_j' iff X_i and X_j stand in the relation G

and so on for each predicate.

2. for all X, A: X satisfies '$-A$' iff X does not satisfy 'A'
3. for all X, A, B: X satisfies '$A \ \& \ B$' iff X satisfies A and X satisfies B
4. for all X, A, i: X satisfies '$(\exists x_i)\ A$' iff there is a sequence Y such that $X_j = Y_j$ for all $j \neq i$ and Y satisfies 'A'

(Notice how each clause of the definition of satisfaction corresponds to a clause in the definition of a wff. This is why it is so convenient to work with minimal vocabulary.) A closed sentence is a wff with no free variables; closed sentences will be satisfied either by all sequences or by none.

Definition of 'true': a closed sentence of O is true iff it is satisfied by all sequences.

Tarski shows that his definition is both materially adequate and formally correct. He also shows that it follows from his definition of

truth that of each pair consisting of a closed sentence and its negation one, and only one, is true. This was to be expected in view of the fact, already observed, that the material adequacy condition rules out non-bivalent theories of truth.

6 Commentary on the semantic theory

Tarski's theory has the distinction of having been criticised both for saying too little:

> the neutrality of Tarski's definition[1] with respect to the competing philosophical theories of truth is sufficient to demonstrate its lack of philosophical relevance. (Black 1948 p. 260)

and for saying too much:

> The Tarskian theory...belongs to factual rather than conceptual analysis...Tarski's theory has plenty of meat to it, whereas a correct conceptual analysis of truth has very little. (Mackie 1973 p. 40)

The question of the philosophical significance of Tarski's theory is evidently a hard one; I shall tackle it in three stages: first by discussing Tarski's own estimate of his theory's significance, and then by discussing the use made of the theory by two writers – Popper and Davidson – who have more ambitious hopes of it than Tarski himself.

(a) *Tarski's own estimate*

Tarski expresses the hope (1944 pp. 53–4) that his definition will do justice to the Aristotelian conception of truth, but sees little point in the question whether that is the 'correct' concept, offering, indeed, to use the word 'frue' rather than 'true' should the decision go against him on that issue (p. 68).

Tarski is also modest about the epistemological pretensions of his theory; he doesn't really understand, he says, what the 'philosophical problem of truth' might be (p. 70), but anyway:

> we may accept the semantic conception[2] of truth without giving up any epistemological attitude we may have had,

[1] Here Black apparently confuses the material adequacy condition with the definition, though elsewhere in the same paper he makes the distinction clearly enough.

[2] The context suggests that Tarski is here concerned primarily with his material adequacy condition.

> we may remain naive realists or idealists, empiricists or metaphysicians... The semantic conception is completely neutral toward all these issues. (p. 71)

Field suggests (1972) that Tarski may have attached metaphysical importance to the constraint on which he insisted (but cf. p. 108n), that truth be defined without the use of semantic primitives: a constraint he justified (1931 pp. 152–3) by urging the superior clarity of syntactic notions. A comment in another paper, 'The establishment of scientific semantics', suggests that he may also have had a deeper significance in mind; after repeating that the use of semantic primitives would threaten clarity, he goes on:

> this method would arouse certain doubts from a general philosophical point of view. It seems to me that it would then be difficult to bring this method into harmony with the postulates of the unity of science and of physicalism (since the concepts of semantics would be neither logical nor physical concepts). (1936 p. 406)

Field's conjecture is that Tarski's intention was to bring semantics into line with the demands of physicalism, the thesis that there is nothing but physical bodies and their properties and relations; and that this is to be achieved by *defining* such non-physical concepts as truth and satisfaction. It is confirmed by a passage, 1944 pp. 72–4, where Tarski defends the semantic conception of truth against the criticism that semantics involves metaphysical elements, by stressing that his definition uses as primitives only logical terms, expressions of the object language, and names of those expressions. The further question, whether Tarski's theory indeed has this significance, is tricky. Field believes that Tarski does not really succeed in reducing semantics to physicalistically acceptable primitives. Tarski defines satisfaction for complex open sentences recursively, in terms of satisfaction for atomic open sentences, but he defines satisfaction for atomic open sentences *enumeratively*, a clause for each primitive predicate of the object language (as it were, 'X satisfies 'x_i is a city' iff X_i is a city, X satisfies 'x_i is north of x_j' iff X_i is north of X_j...' and so forth). Since Field holds that a successful reduction requires more than extensional equivalence of *definiens* and *definiendum*, which is all Tarski's definition guarantees, he finds that Tarski does not, as he hoped, vindicate physicalism. It seems worth observing that there is

a strong tendency for physicalists to be extensionalists, and some reason, therefore, to suppose that Tarski would have thought extensional equivalence a sufficient constraint. The question remains, of course, whether extensional equivalence really is a sufficient constraint upon reductions, or whether, as Field suggests, some stronger requirement is proper.

(b) *Popper's claims on behalf of Tarski's theory*

Popper welcomes Tarski's theory as having:

> rehabilitated the correspondence theory of absolute or objective truth... He vindicated the free use of the intuitive idea of truth as correspondence to the facts... (1960 p. 224)

and he uses Tarski's ideas in developing his own account of the role of truth as a regulative ideal of scientific inquiry.[1]

Is Tarski's a correspondence theory?

According to Popper, Tarski has supplied just what was lacking with the traditional correspondence theories – a precise sense for 'corresponds' (1960 p. 223, 1972 p. 320). Initially, at least, this is puzzling, for Tarski explicitly comments that the correspondence theory is unsatisfactory (1944 p. 54), and observes that he was 'by no means surprised' to learn that, in a survey carried out by Ness, only 15% agreed that truth is correspondence with reality, while 90% agreed that 'It is snowing' is true if and only if it is snowing (1944 p. 70; and see Ness 1938).

So what is it that leads Popper to think of Tarski as having vindicated the correspondence theory? Some comments (e.g. 1960 p. 224) suggest that what he specifically has in mind is Tarski's insistence on the need for a metalanguage in which one can both refer to expressions of the object language and say what the object language says. It is as if he thinks of the left-hand side of each instance of the (T) schema, such as:

'Snow is white' is true iff snow is white

as referring to the language, and the right-hand side to the facts. But this seems a pretty inadequate reason for taking Tarski's to be a correspondence theory, for the material adequacy condition, though

[1] This section amplifies and modifies some points from Haack 1976d; and cf. Sellars 1967 ch. 6 for some pertinent discussion.

its role is to rule out some definitions, certainly does not single out the correspondence theory as uniquely correct; presumably it permits, for instance, a redundancy definition such as Mackie's:

(p) (the statement that p is true iff p)

It is just for this reason that Tarski himself stresses the epistemological neutrality of the (T) schema.

However, though Popper does not explicitly refer to them, there are features of Tarski's *definition* of truth which are reminiscent of correspondence theories. A difficulty here is that it isn't very clear what is required for Tarski's to count as really a version of the correspondence theory; and it is aggravated by Popper's insistence that until Tarski there had been no genuine, no satisfactory, correspondence theory. Still, one can make some progress by comparing Tarski's definition first with the logical atomist version given by Russell and Wittgenstein, and then with Austin's version.

Tarski defines truth in terms of satisfaction, and satisfaction is a relation between open sentences and sequences of objects; the account of satisfaction bears some analogy to Wittgenstein's view of truth as consisting in the correspondence between the arrangement of names in a proposition and the arrangement of objects in the world. On the other hand, Tarski's definition of *truth* makes no appeal to specific sequences of objects, for true sentences are satisfied by all sequences, and false sentences by none. It is symptomatic that logical as well as factual truth is embraced by Tarski's definition; it is surely less plausible to suppose that logical truth consists in correspondence to the facts than that 'factual' truth does. Two historical observations seem called for here: first, that Wittgenstein thought that quantified wffs could be understood as conjunctions/disjunctions of atomic propositions, and that if this were indeed so, Tarski's detour through satisfaction would be unnecessary; and secondly, that Russell allowed 'logical facts'.

Tarski's exploitation of the structure of sentences in the recursive definition of satisfaction is, then, an analogy with Russell's and Wittgenstein's gloss on 'corresponds'. It is equally a *dis*analogy with Austin's account. Austin insists that statements, not sentences, are the primary truth-bearers. This has at least two relevant consequences: Tarski ignores problems raised by sentences containing indexicals such as 'I' and 'now', upon which Austin concentrates; and while Tarski's definition of satisfaction relies on the syntactic structure of open sentences, Austin's account of correspondence stresses its

purely conventional, arbitrary character – in another language, the statement that nuts could, he says, be true in just the circumstance that the statement in English that the National Liberals are the people's choice is true.[1] There is, however, one point of analogy which deserves mention. Austin's account, I suggested earlier, avoids locating correspondence in the too intimate connection between the statement that p and the fact that p, explaining it rather as consisting in the situation to which the statement that p refers being of the kind which the statement says it is. Here one can see, without too severe a strain, a resemblance to Tarski's enumerative account of satisfaction for atomic open sentences: for example, X satisfies 'x_i is white' iff the ith thing in the sequence X is white.

So: Tarski does not regard himself as giving a version of the correspondence theory, and his material adequacy condition is neutral between correspondence and other definitions. However, Tarski's definition of satisfaction, if not of truth, bears some analogy to correspondence theories: the clauses for atomic open sentences to Austin's, the clauses for molecular open sentences to Russell's and Wittgenstein's version.

Is Tarski's theory 'absolute' and 'objective'?

Whether or not one considers the affinities strong enough for Tarski's to count as a version of the correspondence theory, it is worth asking whether the semantic definition of truth has, anyway, what Popper considers to be the major virtues of the correspondence theory, its 'absolute' and 'objective' character.

Tarski stresses that truth can be defined only *relatively to a language* – what he defines is not 'true' (period), but 'true-in-O'. This is for two reasons; that the definition must apply to sentences (which, unlike such extralinguistic items as propositions, have the syntactic structure which it exploits) and one and the same sentence can be true in one language and false, or meaningless, in another; and that only a hierarchy of object language, metalanguage, and meta-metalanguage can avoid the semantic paradoxes. In this sense, therefore, Tarski's is not an absolute, but a relative, truth definition.

[1] Here, then, is a case where the issue about truth-bearers acquires a real significance. (I shall not resist the temptation to draw attention to Austin's complaint (1950 p. 30), that his fellow-symposiast, Strawson, had failed to make the crucial distinction between sentence and statement.) I shall touch on the question, how Taraki's theory might be adapted to deal with indexical sentences, in the section on Davidson.

Popper, however, who is apt to take a somewhat cavalier attitude to the question of truth-bearers (1972 pp. 11, 45, 319n) is unconcerned with this sense of 'absolute'. He takes no interest, either, in the fact that Tarski's original definition is absolute rather than model-theoretic.

Popper seems, rather, to equate 'absolute' and 'objective', contrasting both with 'subjective', that is, 'relative to our knowledge or belief'. In this respect Popper believes the correspondence theory to be superior to

> the coherence theory...[which] mistakes consistency for truth, the evidence theory...[which] mistakes 'known to be true' for 'true', and the pragmatist or instrumentalist theory [which] mistakes usefulness for truth. (1960 p. 225)

I needn't comment, I think, on the accuracy of Popper's characterisation of the rival theories; anyhow, the core of his argument, fortunately, does not depend on these details. The rival theories, Popper argues, are founded on the 'widespread but mistaken dogma that a satisfactory theory should yield a criterion of true belief' (1960 p. 225). And a criterial theory of truth is subjective because it cannot allow the possibility of a proposition's being true even though no one believes it, or false even though everyone believes it.

What exactly does Popper find objectionable about criterial truth-theories? Popper doesn't make this very clear; but I think the problem can be focussed. The crucial difficulty lies, not in the attempt to supply a criterion of truth, in itself, but in the adoption of a criterial theory of the meaning of 'true'. (His attitude is perhaps clearest in the appendix to the 1961 edition of vol. 2 of *The Open Society and its Enemies*.) If one gives the meaning of 'true' in terms of our criteria of truth, one cannot leave room for the possibility that a proposition be false though it passes our tests of truth, or true though it fails them. This is a particular problem for the pragmatists, since it poses a threat to their official fallibilism; though there is still room for mistakes in the *application* even of infallible tests of truth. Infallibilism *in itself* is not subjectivist; but the further claim that to say that a proposition is true (false) *just means* that it passes (fails) our tests poses a threat to objectivism.

Tarski expressly disclaims the aspiration to supply a criterion of truth (1944 pp. 71–2); and his definition certainly makes no reference to our tests of truth. (Ironically, the passage in which Tarski draws

attention to these features is intended as a rebuttal of the 'objection' that his is a kind of correspondence theory which involves logic in 'a most uncritical realism'!)

So Tarski's is an objective theory in Popper's sense. But why does Popper attach so much importance to this point? The explanation lies in the epistemological use to which he proposes to put the concept of truth.

Truth as a regulative ideal: verisimilitude

Popper describes himself as a 'fallibilistic absolutist': fallibilist because he denies that we have any guaranteed method of acquiring knowledge; absolutist because he insists that there is such a thing as an objective truth to which scientific inquiry aspires. Tarski's theory is to supply a suitably objective account of this 'regulative ideal' of science.

This requires, of course, that Tarski's theory be applicable to the languages – presumably, fragments, more or less completely regimented, of natural and mathematical languages – in which scientific theories are expressed. I shall not here discuss the questions raised by this requirement; partly because Tarski himself expresses (1944 p. 74) a cautious optimism about the applicability of his work to the empirical sciences, and partly because in the next section, when I discuss Davidson's use of Tarski's work, I shall have to consider the reasons Tarski gives for doubting whether his methods apply to 'colloquial' language.

According to Popper, the business of science is to devise and test conjectures; scientists can't be confident that their current conjectures are true, nor even that they will ever reach the truth or would know, if they did reach the truth, that they had. But if truth is to be not just an ideal, but a *guiding* or 'regulative' ideal, it should be possible to tell whether, as one theory replaces another, science is getting closer to the truth. So Popper's problem is to explain in what sense, of two theories both of which may be false, one can be closer to the truth than another. His solution is his extension of Tarski's ideas in the theory of 'verisimilitude', or truth-likeness.

Popper's account of verisimilitude goes:

> *Assuming that the truth-content and the falsity-content of two theories t_1 and t_2 are comparable, we can say that t_2 is more closely similar to the truth...than t_1 if and only if either:*

> (a) *the truth-content but not the falsity-content of t_2 exceeds that of t_1*
> [or]
> (b) *the falsity-content of t_1 but not its truth-content exceeds that of t_2.* (1963 p. 233)

The truth- (falsity-) content of a theory is the class of all and only its true (false) consequences. The truth- or falsity-content of one theory can exceed the truth- or falsity-content of another only if its truth- or falsity-content set-theoretically includes the other's, so this account applies only to theories which overlap in this way. Popper also suggests (1963 pp. 393–6, 1972 pp. 51, 334) measures of truth- and falsity-contents in terms of logical probability, so that any two contents can be compared. But I shall concentrate on the former, 'qualitative', rather than the latter, 'quantitative' version.

The definition of verisimilitude cannot show that science does progress towards the truth: but Popper hopes (1972 p. 53) that it supports his falsificationist methodology, which recommends that one choose the more falsifiable conjecture, the one with more content, for a theory with more content will have greater verisimilitude, unless, Popper adds, it has more falsity-content as well as more truth-content.

However, it has been shown[1] that a theory t_2 has greater verisimilitude than another, t_1, in accordance with Popper's (a) and (b) only if t_2 is a *true* theory from which the truth-content of t_1 follows. This means that *Popper's definition of verisimilitude does not apply to comparisons between theories both of which are false*; but that, of course, was the principal objective of the theory, which therefore fails of its epistemological purpose. This failure is, I think, important to the question of the feasibility of fallibilistic absolutism (see Haack 1977b); and it should also, to my mind, support Tarski's rather modest, as against Popper's rather more ambitious, assessment of the epistemological significance of the semantic theory of truth.

[1] Miller 1974; and cf. Tichý 1974 and Harris 1974. Very briefly, Miller's strategy is first to show, if t_1 and t_2 are comparable by truth-content, how they are also comparable by falsity-content; and then to show that for t_2 to be nearer the truth than t_1, t_2 must be a true theory from which the truth-content of t_1 follows, since otherwise t_2 will exceed t_1 in falsity- as well as truth-content, so that their verisimilitudes will not be comparable.

(c) *Davidson's use of Tarski's theory*

Truth and meaning. Any adequate theory of meaning, Davidson thinks, must explain how the meanings of sentences depend upon the meanings of words (otherwise, he argues, the language would be unlearnable). A theory of meaning must be consistent with – or, he sometimes says, explain – 'semantic productivity': speakers' ability to produce, and understand, sentences they have never heard before. What this amounts to, he claims, is that the theory should yield all sentences of the form:

S means m

where 'S' is a structure-revealing description of a sentence of the language for which the theory is being given, and 'm' a term denoting the meaning of that sentence. But the appeal to meanings implicit here, he suggests, contributes nothing useful; and reformulating the requirement thus:

S means that p

where 'p' is a sentence that has the meaning that the sentence described by 'S' has, leaves a problem with the 'means that', which, therefore, Davidson reformulates as 'is T iff' where 'T' is any arbitrary predicate which, given the above conditions on 'S' and 'p', satisfies:

S is T iff p

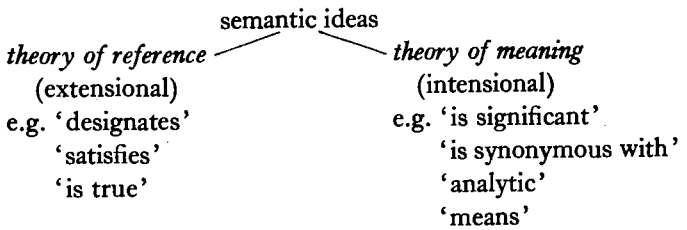
But, of course, any predicate satisfying this condition will be, by Tarski's standards, a materially adequate *truth*-predicate. Davidson concludes that what is required by a theory of meaning is, precisely, a definition of such a truth-predicate (Davidson 1967).

Meaning as truth-conditions

Though the route by which Davidson reaches this conclusion is somewhat indirect, the terminus – that the meaning of a sentence can be given by specifying its truth-conditions – is not unfamiliar. What is novel in Davidson's version is the imposition of 'Tarskian' constraints upon the account of truth-conditions.[1]

[1] Dummett urges a theory of meaning in terms rather of assertibility-conditions than truth-conditions (again a comparison with the pragmatists, now with their criterial theory of meaning, suggests itself). For critical discussions see Haack 1974 pp. 103ff. and cf. Brandom 1976.

The appeal of a truth-condition theory of meaning may perhaps be appreciated by recalling Quine's classification of semantic notions into two groups, the extensional, which he takes to constitute the business of the 'theory of reference' and the intensional, which he takes to constitute the business of the 'theory of meaning', thus:



Quine argued in 1953a that the theory of reference was in considerably better shape than the theory of meaning. An appealing feature of the truth-condition theory is that it promises an explanation of meaning (from the more problematic right-hand side) in terms of truth (from the less problematic left-hand side).

Theory of interpretation

Later (1974) Davidson appends some further theory, of interpretation of another's discourse, in another or even in the same language as one's own; essentially, this consists in an account of how to tell when 'p' is a sentence that has the meaning that 'S' describes. Briefly, the idea is that to test, empirically, whether a sentence of the form

'*Es regnet*' is true iff it is raining

is a T-sentence, that is, meets Tarski's specification that the sentence on the right translates the one named on the left, one tests whether speakers of the language concerned (here, German) hold true '*Es regnet*' iff it is raining. The point of the appeal to what native speakers *hold true* is to get at the meaning of their utterances by, so to speak, holding their beliefs constant. In consequence an assumption, the principle of charity, to the effect that speakers of other languages generally agree with us about what is the case, is required. The holist character of Davidson's account, the insistence that the 'unit of interpretation' is the entire language, may derive at least in part from epistemic holism, the Duhemian idea, also stressed by Quine, that beliefs are verified/falsified not alone but in a corporate body.

Though there are many important questions to be asked about this theory of interpretation, I shall concentrate, in what follows, on Davidson's account of meaning, since it is there that Tarski's theory of truth plays the crucial role.

If the task of a theory of meaning is indeed, as Davidson thinks, to define a Tarskian truth-predicate, what work over and above that already accomplished by Tarski would be needed? Davidson is seeking a theory of meaning *for natural languages*, such as English; Tarski, of course, is thoroughly sceptical about the applicability of his theory to natural languages. So a first task, if Davidson's programme is to be feasible, is to show that Tarski's methods *can* be extended. This is an important question even independently of Davidson's special ambitions for Tarski's methods, for the concept of truth is of philosophical significance in many contexts where 'true' must be allowed to apply to sentences of natural languages – in epistemology, for instance. Despite Tarski's official modesty on this score, it seems to me that the usefulness of his work would be sadly restricted if the concept he defines turns out to be quite different from the concept of truth in natural languages.

Is Tarski's theory applicable to natural languages?
According to Tarski:

> *The very possibility of a consistent use of the expression 'true sentence' which is in harmony with the laws of logic and the spirit of everyday language seems to be very questionable, and consequently the same doubt attaches to the possibility of constructing a correct definition of this expression.*
> (1931 p. 165)

Tarski's pessimism has two main sources: his formal correctness condition rules out the possibility of an adequate definition of truth for languages which are neither (i) semantically open nor (ii) formally specifiable. Natural languages, Tarski argues, fail on both scores, so there is no prospect of an adequate definition of truth for them.

(i) Tarski suggests that natural languages contain their own metalanguages, so that truth cannot be defined without running into paradox; though sometimes he hints, rather, that because natural languages are not formally specifiable, the question of their semantical closure cannot be answered. Davidson has no very satisfactory answer to this problem, but urges that 'we are justified in carrying on without having disinfected this source of conceptual anxiety' (1967 p. 10). He seems to propose that work proceed on those semantically open fragments of natural languages where the danger of paradox does not arise. There is some difficulty in squaring Davidson's attitude to the paradoxes (don't worry too much about them, concentrate on the rest of the job) with his holism, the insistence that an adequate theory of meaning must be a theory for a whole language; though he also hints that he doubts whether natural languages really are universal.

(ii) There seems to be a whole family of difficulties here; the problem of giving a precise account of just what strings count as sentences of a natural language, aggravated by the fact that natural languages are not static, but growing; and the prevalence in natural languages of such phenomena as vagueness, ambiguity, indexicality. Tarski is gloomy:

> Whoever wishes, in spite of all difficulties, to pursue the semantics of colloquial language with the help of exact methods will be driven first to undertake the thankless task of a reform of this language... It may however be doubted whether the language of everyday life, after being 'rationalised' in this way, would still preserve its naturalness and whether it would not rather take on the characteristic features of the formalised languages.
> (1931 p. 267)

The core of Davidson's reply to this is that, though some 'tidying up' will be needed before Tarski's methods can be applied to a natural language, this need not be such as to transform it out of all recognition. He would hold, I think, that work in transformational grammar (see e.g. Chomsky 1957) promises to overcome the first problem; and he is optimistic that more fragments of natural languages can be brought within the scope of Tarskian methods, rather as Frege's work on '(x)' and '$(\exists x)$' has already suitably regimented 'all', 'none' and 'some'.[1]

What Tarski regards as a 'thankless task' Davidson undertakes gladly, observing that 'It's good to know we shan't run out of work'. His main task, in fact, is to supply a suitable analysis of those locutions

[1] It is not beyond dispute that Frege's account does properly regiment natural language quantifiers; recall (ch. 4 § 1) that Montague and Hintikka, like the early Russell, stress their affinities with singular terms, whereas according to Frege they belong to an entirely different syntactic category.

of natural languages which are initially recalcitrant to Tarskian treatment. And it is on his success or failure at this task that one's assessment of Davidson's response to Tarski's scepticism must be based. It is worth observing that Davidson insists on using the 'absolute' rather than a model-theoretic concept of truth; and that some of these problems (e.g. problems created by the introduction of new predicates as a natural language grows) are harder on an absolute than they would have been on a model-theoretic approach; cf. Field 1972.

Logical form

Davidson describes himself as seeking 'the logical form' of natural language locutions. For example, recall (ch. 2 §4) that, according to Davidson, adverbial constructions in natural language are best represented as involving quantification over events, with adverbs construed as adjectives of event terms. The logical form of 'John buttered the toast with a knife', Davidson claims, is something like 'There is an event which is a buttering of the toast by John and which is performed with a knife'. Davidson's confidence that each natural language construction has a unique logical form springs from the belief that a formal representation to which Tarski's method of defining truth applies represents essential structure in an ideally perspicuous fashion. (The analogy with Russell's and Wittgenstein's project, in their logical atomist periods, of devising an ideal language which would represent the *real* form of natural languages, is striking.) Interestingly, Cargile has asked (1970; and cf. Davidson's reply in the same volume) why the connection between a predicate and its adverbially modified form need necessarily be assumed to be a matter of form rather than content. It isn't, he suggests, as obvious as Davidson seems to suppose what one should count as skeleton, and what as flesh; he urges, in fact, a more flexible conception of logical form, closer to the one presented in ch. 2.

Davidson's programme

So Davidson regards it as the task of a theory of meaning to analyse the structure of sentences, not to supply an account of the meaning of individual words. (This isn't *quite* right, because some particles – 'un' for instance – have a structural character.) For example, Davidson does not require a theory of meaning to give the meaning of 'good', but he does require it to analyse the structure of e.g. 'Bardot is a good actress' in such a way as to explain why it is not equivalent

to 'Bardot is good and Bardot is an actress' as 'Bardot is a French actress' is to 'Bardot is French and Bardot is an actress' (cf. 'small elephant', and the ambiguous 'poor violinist'). The appeal of Tarski's method, which is to define satisfaction for complex open sentences in terms of satisfaction of simple open sentences, is its promise of an explanation of how the meanings of compound sentences depend on the meaning of their parts; the challenge is to analyse sentences like 'Bardot is a good actress' so that Tarski's method applies to them as well as to the less recalcitrant 'Bardot is a French actress'. Davidson admits that the task is considerable, that:

> a staggering list of difficulties and conundrums remains. (1967 p. 321)

He includes ('to name a few') counterfactuals, subjunctives, probability statements, causal statements, adverbs, attributive adjectives, mass terms, verbs of belief, perception, intention, action. Obviously my consideration of details of the programme will have to be selective.

Indexicals

Tarski's theory needs to be relativised to speakers and times, Davidson suggests, because natural languages contain indexicals. The revised (T) schema will call for the theory to entail sentences like:

> 'I am tired' (s, t) is true iff s is tired at t

Truth, Davidson says, is a predicate rather of utterances than sentences. (This suggestion is relevant to the claim, canvassed by Strawson and, before him, by Schiller, that formal methods are inherently inadequate to deal with the context-dependence of statements in natural languages.)

But Davidson's concern with indexicals is also directed towards the problems raised by the analysis of quotation and verbs of 'propositional attitude' ('says that', 'knows that' etc.); for he thinks that these constructions all involve concealed demonstratives. An analysis of these indexicals ('this', 'that') given by Weinstein 1974 has been endorsed by Davidson. In this account, 'That is a cat', say, is true just in case the object indicated by the speaker at the time of utterance satisfies '...is a cat'.

Oratio obliqua

While truth-functional compounds raise no problems, there will obviously be a difficulty about applying Tarski's methods to compound English sentences the truth-values of which do not depend in any obvious way upon the truth-values of their parts. *Oratio obliqua* sentences are of this problematic, intensional kind; for the truth-value of 'Galileo said that the earth moves', for instance, does not depend in any direct way on the truth-value of 'the earth moves'; and there is failure of substitutivity, for from 'Tom said that the moon is round' and 'The moon = the sole planet of the earth' one cannot safely infer 'Tom said that the sole planet of the earth is round'.

The first step in the right direction, Davidson urges, is to parse:

> Galileo said that the earth moves.

along the lines of:

> Galileo said that.
> The earth moves.

The 'that' is to be construed not as a relative pronoun, but as a demonstrative pronoun referring to an utterance – rather as I might say 'I wrote that', pointing to a message on the notice-board.[1] Of course, Galileo didn't utter the very utterance which the speaker produces; indeed, Galileo didn't speak English; so some more explanation is needed. Davidson amplifies his analysis thus:

> The earth moves.
> $(\exists x)$ (Galileo's utterance x and my last utterance make us samesayers)

Galileo and I are samesayers, we are told, just in case he uttered a

[1] Two points of comparison are worth making. I have already mentioned the affinities between Tarski's definition of satisfaction and Wittgenstein's *Tractatus* account of truth; the verbs of propositional attitude, which present a problem to Wittgenstein's as to Davidson's approach, are discussed at 5.542. Wittgenstein's analysis is, however, notoriously obscure. Alun Jones has pointed out to me that Davidson's list of 'difficulties and conundrums' for *his* enterprise, and Anscombe's list (1959) of the problems for Wittgenstein's, are very similar. An analysis of indirect discourse strikingly like Davidson's was suggested by Kotarbiński (1955). Kotarbiński's aim was to support the thesis that only material bodies exist ('pansomatism') by analysing away apparent references to such abstract objects as propositions; this, in view of the conjecture that Tarski was motivated by sympathy with materialism, may be significant.

sentence which meant in his mouth what some utterance of mine meant in my mouth.

The application of Tarski's methods, as extended by Weinstein to cope with indexicals, gives a result along the lines of:

> 'Galileo said that the earth moves' (s, t) means that / is true iff
> Galileo uttered at t' (t' earlier than t) a sentence which meant in his mouth what the utterance demonstrated by s at t'' (t'' just after t) meant in s's mouth, where the demonstrated utterance is of 'The earth moves'.

It may be useful to pause briefly to contrast Davidson's with some alternative accounts of the (so-called) 'propositional' attitudes. Frege, for instance, would regard 'that p' in 's said (believes) that p' as referring to a proposition (see ch. 5 §2). Carnap would analyse 's said (believes) that p' as 's uttered (is disposed to assent to) some sentence intensionally isomorphic to 'p' in English' (see 1947). Scheffler treats 'that p' rather as an adjective than a noun: 's said (believes) that p' amounts to 's uttered (believes) a that-p utterance' where there is a separate predicate corresponding to each sentence 'p' (see 1954); Quine goes yet further in the same direction, treating the whole of 'said (believes)-that-p' as a predicate of s (see 1960a §44).

Davidson believes his account to have the advantages that: unlike analyses which treat 'that p' as referring to a proposition, it doesn't require appeal to intensional entities; unlike Carnap's analysis, it does not require explicit reference to a language; and unlike analyses which treat 'says (believes)-that-p' as a single predicate, it allows that what follows the 'that' is a sentence with 'significant structure', structure a theory of meaning can exploit.

His account, Davidson argues, allows, as seems proper, that 's said that p' entails 's said something', for the analysis goes 's uttered a sentence which...'. At the same time, it explains, as is also required, why 's said that p' does not entail 'p', for what seemed to be a sentence ('p') within a sentential operator ('s said that') becomes a single sentence ('s said that') containing a demonstrative ('that') which refers to an utterance of another sentence ('p'). And just as, although cats scratch, the sentence, 'that is a cat', which refers to a cat, doesn't scratch, so, although 'p' entails 'p', the sentence 's said that p', which refers to an utterance of 'p', doesn't entail 'p'.

Of course, as the last example brings out, in the regular cases considered by Weinstein, what 'that' refers to is a non-linguistic item, a cat, for instance. When the account is extended to 'that's in indirect speech, the referents will be utterances of sentences. And these sentences have significant structure (among the instances of '*s* said that *p*' would be e.g. '*s* said that *q* and *r*' and '*s* said that *s'* said that *q*') in virtue of which their meaning would be given.

Some comments are in order, here, about how Davidson's differs from Carnap's analysis. On Davidson's account '*s* said that *p*' involves reference to an utterance of the speaker's related to some utterance of *s*'s by *samesaying*; in Carnap's, reference to a sentence related to a sentence of English by *intensional isomorphism*. An utterance (here, Davidson makes it clear that he means a speech act, the event of uttering a sentence) is an utterance of some sentence in some specific language with some specific context; and so the need to specify the relevant language is avoided.[1]

This gives Davidson's account an unexpected character – for the concept of utterance (speech act) belongs rather to pragmatics than to semantics. Equally surprising, and methodologically also disquieting, is that Davidson's account, like Carnap's, requires a semantic primitive (respectively, samesaying and intensional isomorphism) in the metalanguage. *s* and *s'* are samesayers, Davidson explains, just in case some utterance of *s*'s *means the same as* some utterance of *s'*'s. Now, Davidson insists that the truth-conditions be given in terms of an absolute definition of truth, a definition, that is, which uses no semantic primitives. And he avoids '*S* means *m*' and the formula '*S* means that *p*' because of their intensional character. Davidson apparently regards the appeal to samesaying as admissible because *local*; the general account of meaning appeals only to Tarskian truth-conditions, though the specific account of 'says that' requires samesaying as semantic primitive. It is questionable, though, whether the appeal *is* local in the relevant sense; for surely 'says that' counts as structure rather than vocabulary in the sense in which the dependence of the meaning of 'good' on the meaning of 'actress' in 'Bardot is a good actress' is structural (Davidson objects to the Fregean account *of indirect discourse* because it requires intensional objects). The problem is what exactly the constraints should be on Davidson's enterprise: what apparatus should he be permitted to use, and where? It is pertinent that the appeal of his enterprise derives in large part from the austerity of method it appears, at the outset, to promise.

[1] Davidson sometimes speaks as if it is the reference to an utterance (rather than a sentence) that prevents '*s* said that *p*' entailing '*p*'. But this is surely sufficiently explained by appeal to the fact that (an utterance of) '*p*' is *referred to by*, and not *contained in* '*r* said that *p*'. The sense of 'utterance' in which, according to Davidson, truth is a property of utterances, has, presumably, to be the 'content', and not, as in this context, the 'act' sense.

Since the enterprise was launched, Davidson and his followers have tackled, with various degrees of success, many of the 'difficulties and conundrums' pointed out in 1967. By 1973 Davidson speaks of 'fairly impressive progress', pointing to work on propositional attitudes, adverbs, quotation (Davidson 1967, 1968a, b), proper names (Burge 1973), 'ought' (Harman 1975), mass terms and comparatives (Wallace 1970, 1972).

The success of Davidson's programme would vindicate, in large measure, the applicability of Tarski's theory to natural languages; but the assessment of his programme obviously depends on the detailed study of the specific analyses offered. And as I have suggested with reference to the analysis of *oratio obliqua*, this study in its turn raises some methodological questions which are at any rate tricky enough that one cannot say with any confidence that Davidson *has* shown that Tarski's theory applies to English.

7 The redundancy theory

Ramsey

The redundancy theory (though suggested earlier by some remarks of Frege in 1918) derives primarily from the work of F. P. Ramsey in 1927. Ramsey offers his sketch of a theory in a very brief passage (pp. 142–3) in the course of a discussion of the proper analysis of belief and judgment; the context is significant of Ramsey's estimate of the importance of the issue: 'there is' he thinks, 'really no separate problem of truth, but merely a linguistic muddle'.

Briefly, his idea is that the predicates 'true' and 'false' are redundant in the sense that they can be eliminated from all contexts *without semantic loss*;[1] he allows that they have a pragmatic role, for 'emphasis or stylistic reasons'. Ramsey considers two kinds of case where 'true' and 'false' typically occur. The cases he uses to introduce the theory are of the more straightforward kind, where the

[1] There is an allusion here to Russell's doctrine of 'incomplete symbols', symbols, that is, which are contextually eliminable. Cf. ch. 5 §3 for a discussion of this doctrine with reference to Russell's theory of descriptions.

proposition to which truth or falsity is ascribed is explicitly given: 'it is true that p', Ramsey argues, *means the same* as 'p', and 'It is false that p' *means the same* as 'not p'. Cases where the relevant proposition is not actually supplied but only described present rather more initial difficulty, for, as Ramsey realises, one cannot simply eliminate 'is true' from, for instance, 'what he says is always true'; this difficulty he proposes to overcome by using the apparatus of propositional quantification, to give, in the case mentioned, something along the lines of 'For all p, if he asserts p, then p'.[1]

Whether the second-order quantifiers which Ramsey needs can be suitably explicated is a key question, as it turns out, for the feasibility of the redundancy theory; but I shall begin by pointing out some of the advantages of the theory before turning to its problems.

Truth-bearers

In view of the embarrassments caused by the trappings – facts and propositions – of the correspondence theory the austerity of the redundancy theory is appealing. Ramsey understandably regards it as a virtue of his theory that it avoids the questions raised by a correspondence account about the nature and individuation of facts. 'It is a fact that...', he urges, has the same semantic redundancy, and the same emphatic use, as 'It is true that...'.

Again, since the effect of Ramsey-style theories is to deny that in 'It is true that p', '...is true...' is to be thought of as a predicate ascribing a *bona fide* property to whatever 'p' stands for, the question of the truth-bearers is similarly bypassed; if truth isn't a property, one needn't ask what it's a property *of*. I observe, however, that what I argued (ch. 6 §5) to be the real issue lying behind disputes about truth-bearers – the question of the appropriate constraints on instances of sentence letters, i.e. what one can put for 'p' – does still arise. (Ramsey's preference for the locution 'It is true that p', rather than ''p' is true' is of some significance in this regard.) I should count it an advantage of my diagnosis of the issue about truth-bearers that it is applicable even to redundancy theories, and an advantage

[1] Tarski writes (1944 pp. 68–9) as if Ramsey's theory simply has no way to handle this kind of case; Ramsey would presumably analyse the two problematic cases Tarski gives – 'The first sentence written by Plato is true' and 'All consequences of true sentences are true' – as '(p) (if the first thing Plato wrote was that p, then p)' and '(p) (q) (if p, and if p then q, then q)'.

of the redundancy theory that there the issue arises in its fundamental form.

Of course, this will be a genuine economy only if it is certain that one doesn't need propositions (or whatever) for other purposes besides truth-bearing. Those who believe that we need propositions as objects of belief, for instance, are liable to be less impressed by the redundancy theory's ability to do without them as bearers of truth. It is significant, therefore, that Prior, who accepts Ramsey's theory, urges (1971 ch. 9) an account of belief according to which 's believes that...' in 's believes that p' is a sentence-forming operator on sentences like 'It is not the case that...', rather than 'believes' being a relation symbol with arguments 's' and 'that p', the latter denoting a proposition. Again, one might suppose that propositions (or whatever) may be required as bearers of *other* properties, and that the redundancy theory is therefore in danger of sacrificing the analogy between '...is true', and, say, '...is surprising' or '...is exaggerated' without, in the end, any compensation by way of genuine ontological economy. And it is significant, in this regard, that Grover *et al.*, in a paper (1975) urging the claims of a redundancy-style theory, argue that it is only a misleading appearance that '...is true' and '...is surprising' are ascriptions to the very same thing.

The object language/metalanguage distinction

The redundancy theorist denies that 'It is true that p' is about the sentence 'p': 'It is true that lions are timid', like 'It is not the case that lions are timid', is in his view about lions, not about the sentence 'Lions are timid'. This means that he sees no need for insistence on the distinction between object language and metalanguage which is so vital to Tarskian semantics (Prior shows most awareness of this point; e.g. 1971 ch. 7). This raises some questions about the redundancy theory's capacity to handle problems where the object language/metalanguage distinction apparently plays an important role.

The idea that truth is a metalinguistic predicate seems, for example, to contribute to the usual explanations of the semantics of the sentence connectives, as: ''$-p$' is true iff 'p' is false', '$p \vee q$' is true iff 'p' is true or 'q' is true'. How adequate an alternative theory can the redundancy theory offer? Since that theory equates both 'It is false that p' and 'It is true that $-p$' with '$-p$', all that remains of the 'explanation' of negation seems to be '$-p$ iff $-p$'. The redundancy theorist might urge that there is indeed less than meets the eye to the

usual explanations of negation, for there is, according to him, less than meets the eye to the usual explanations of truth. (Cf. Dummett 1958, and Grover *et al.*'s acknowledgment that 'It is not the case that...' may not be eliminable.)

Another, related, difficulty is that the redundancy theorist seems to be unable to allow an apparently genuine distinction between the law of excluded middle ('$p \vee -p$') and the metalinguistic principle of bivalence ('for all p, 'p' is either true or else false'). For if ''p' is true' means the same as 'p', and ''p' is false' means the same as '$-p$', then ''p' is either true or else false' means '$p \vee -p$'. Once again, the redundancy theorist might accept the consequence, that this is a 'distinction' without a difference; but since it is a distinction with, apparently, some explanatory power, this leaves him with some explaining to do. (For instance, would he insist that van Fraassen's 'supervaluational' languages, where '$p \vee -p$' is a theorem but the semantics allow truth-value gaps, must be confused? Cf. Haack 1974 pp. 66 ff. and ch. 11 §4 below.)

I pointed out above (pp. 101–2) that the (T) schema seems to require bivalence, and this raises the question whether a redundancy theory isn't also committed to the thesis that 'p' must be either true or else false. But this consequence is avoidable, for the redundancy theorist may deny that, if it is neither true nor false that p, it is false that it is true that p; after all, since his theory is that 'it is true that p' means the same as 'p', he could reasonably insist that, if it is neither true nor false that p, it is also neither true nor false that it is true that p. So he isn't obliged to deny the possibility of truth-value gaps and hence, the previous argument doesn't entail that he is obliged to insist on the law of excluded middle.

In Tarski's work, of course, the most important role of the object language/metalanguage distinction was to secure formal adequacy, specifically, to avoid the semantic paradoxes. So its capacity to deal with the paradoxes will be a pretty crucial question for one's assessment of the feasibility of the redundancy theory. This question must wait till ch. 8; but some of the considerations about propositional quantifiers, to which I now turn, will be relevant to it.

The quantifiers: '(p) (*if he asserts that* p, p)'

Ramsey proposes to eliminate 'true', where what is said to be true is not explicitly supplied but only obliquely referred to, by means of second-order quantification: 'What he says is always true', for

instance, is to be explained as meaning 'For all p, if he asserts p, then p'. He admits that there is some awkwardness in this analysis, for, he thinks, English idiom seems to call for a final 'is true' (as: '(p) (if he asserts p, then p is true)' to make the final 'p' into a *bona fide* sentence; but this apparent obstacle to elimination is overcome, he argues, if one remembers that 'p' is itself a sentence, and already contains a verb. Supposing that all propositions had the logical form '$a\ R\ b$', he suggests, one could observe the grammatical proprieties by writing 'For all a, R, b, if he asserts $a\ R\ b$, then $a\ R\ b$'. But of course, as Ramsey is well aware, all propositions are *not* of the form '$a\ R\ b$', and neither is there much prospect of giving a finite disjunction of all possible forms of proposition, so this scarcely solves the problem.

Ramsey's discomfort is understandable, for the problem is real. If, in his formula:

(p) (if he asserts p, then p)

the quantifier is interpreted in the standard, objectual style, one has:

For all objects (propositions?) p, if he asserts p, then p

Here the bound 'p's are syntactically like singular terms, and the final 'p' has, therefore, to be understood elliptically, as implicitly containing a predicate, to turn it into something of the category of a sentence, capable of standing to the right of 'then', along the lines of:

For all propositions p, if he asserts p, then p *is true*

But if the analysis turns out to contain the predicate 'is true', truth hasn't, after all, been eliminated, and it isn't, after all, redundant. (This is the difficulty Ramsey sees; it is stated rather clearly, with reference to Carnap's version of the redundancy theory, in Heidelberger 1968.) If, on the other hand, the quantifier is interpreted substitutionally, one has:

All substitution instances of 'If he asserts...then...'
are true

and once again 'true' appears in the analysis, and so, hasn't really been eliminated.

So this much is clear: if Ramsey's theory is to work, some *other* explication of the second-order quantifiers will be needed, since on either of the usual interpretations, 'true' seems not to be eliminated.

Prior sees the difficulty as the result of a deficiency in English, which lacks suitable colloquial locutions for reading second-order quantifiers, and obliges one to resort to such misleadingly nominal-sounding locutions as 'Every*thing* he says...'. He therefore suggests (1971 p. 37) 'anywhether' and 'somewhether' as readings of '(p)' and '$(\exists p)$', and reads '$(p)\,(p \rightarrow p)$', for instance, as 'If anywhether, then thether'.

Grover also thinks that the quantifiers can be supplied with suitable readings, and offers some further grammatical apparatus to this purpose. The difficulty of giving an appropriate reading arises, as Prior suggests, from the lack of words and phrases to stand in for sentences in the way that *pronouns* stand in for names and descriptions; what is needed, as Grover puts it, is *prosentences*.

Pronouns and prosentences are two kinds of *proform*; cf. proverbs like 'do', and proadjectives like 'such'. A proform must be capable of being used anaphorically, for cross-reference, either like pronouns of laziness (Geach 1967) as in '*Mary* meant to come to the party, but *she* was ill', or like 'quantificational' pronouns, as in 'If *any car* overheats, don't buy *it*'. Prosentences are like pronouns in occupying positions that sentences could occupy, as pronouns occupy positions that nouns could occupy, and fulfil a similar anaphoric role. Grover's proposal is that one read '(p) (if he asserts that p, then p)' as:

For all propositions, if he asserts that thatt, then thatt

where 'thatt' is a prosentence. Notice that what is proposed is a novel *reading*; it is, Grover argues, compatible with either an objectual or a substitutional account at the level of formal interpretation.

This ingenious proposal raises a number of questions, to which I can offer only tentative answers. First, remember that the problem with which I began was whether it is possible to give a reading of Ramsey's propositional quantifiers which is grammatical, and which doesn't re-introduce the predicate 'true'. Does Grover's reading meet these requirements? Well, it would be somewhat odd to ask whether her reading is grammatical, since it isn't, of course, English; it expressly calls for an *addition* to English. It would be more appropriate to ask whether there are sufficiently strong grammatical analogies to justify her innovation; but this, in view of the 'sufficiently strong', is none too precise a question. English, as Grover allows, doesn't have any atomic prosentences – though it does, I think, have compound expressions that play such a role: 'It is', for instance,

which one might describe as a prosentence composed of a pronoun and a proverb. And the second part of the question, whether Grover's reading genuinely eliminates 'true', is equally tricky. In fact, there are two points to be raised here. The first is that even if a suitable *reading* is supplied, this leaves open a question about whether there isn't still an implicit appeal to truth at the level of formal interpretation. (And what exactly must one eliminate 'true' *from* to show that it is redundant?) The second question is whether one's understanding of 'thatt' implicitly requires the notion of truth.

The 'prosentential theory of truth'

Some light may be shed on this problem by Grover's own application of her account of propositional quantification to the theory of truth. Grover *et al.* 1975 propose a modified version of the redundancy theory according to which 'that is true' is explained as being itself a prosentence. Truth-ascriptions, on their account, are eliminable in favour of 'It is true' *as an atomic prosentence*, i.e. one in which 'true' is not a separable predicate.[1]

What does this show about whether the 'prosentential' theory really eliminates 'true'? 'True', one is told, is eliminable; not from English, to be sure, but from English + 'thatt'. But how are we to understand 'thatt'? Well, there's nothing *exactly* like it in English, but it works like 'That's true', except for being atomic rather than compound....

It is open to doubt, I think, whether Ramsey's hope of eliminating talk of truth altogether has been vindicated. Nevertheless, there is something important to be learned from the discussion of the prosentential theory: that the truth-predicate plays a crucial role in enabling us to talk *generally*, to talk, that is, about propositions which

[1] Ramsey thought that all truth talk is eliminable; Grover *et al.* admit that there is a residue. In some cases the elimination of 'true' calls for modification of the contained sentence, as 'It used to be true that Rome was the centre of the known world'/'Rome used to be the centre of the known world' or, 'It might be true that there is life on Mars'/'There might be life on Mars'. And where this phenomenon is combined with quantification, as in 'Some sentences used to be true but are true no longer', they are obliged to introduce new connectives, as '$(\exists p)$ (it-used-to-be-the-case-that p but it-is-no-longer-the-case-that p)', which they admit to be, in effect, truth-locutions. Their comments about 'It might be true that', on the other hand, suggest an interesting alternative to the idea that necessary truth, like truth, is a property of sentences or propositions.

we don't actually exhibit, but only refer to indirectly, a role it shares with the apparatus of second-order ('propositional' or 'sentential') quantifiers. This similarity of function will turn out to be relevant to the diagnosis of the semantic paradoxes.

8

Paradoxes

1 The Liar and related paradoxes

The importance of the Liar paradox to the theory of truth has already become apparent; for Tarski's formal adequacy conditions on definitions of truth are motivated, in large part, by the need to avoid it. It is time, now, to give the Liar and related paradoxes some direct attention on their own account.

Why the 'Liar paradox'? Well, the Liar sentence, together with apparently obvious principles about truth, leads, by apparently valid reasoning, to contradiction; that is why it is called a paradox (from the Greek, '*para*' and '*doxa*', 'beyond belief').[1]

The Liar comes in several variants; the classic version concerns the sentence:

(*S*) This sentence is false

Suppose *S* is true; then what it says is the case; so it is false. Suppose, on the other hand, that *S* is false; then what it says is not the case, so it is true. So *S* is true iff *S* is false. Variants include indirectly self-referential sentences, such as:

The next sentence is false. The previous sentence is true.

and the 'postcard paradox', when one supposes that on one side of a postcard is written:

The sentence on the other side of this postcard is false

and on the other:

The sentence on the other side of this postcard is true.

[1] The 'paradoxes' of material and strict implication – discussed at length in ch. 11 – are, at worst, counter-intuitive, and not, like the Liar, contradictory; hence the scare quotes.