PUTNAM,
"The MEANING of MEANING"

Mind, Language and Reality

Philosophical Papers, Volume 2

HILARY PUTNAM
Professor of Philosophy, Harvard University

CAMBRIDGE
UNIVERSITY PRESS

short, the positivist attitude tends to be that social science is science only when and to the extent that it apes *physics*. And this for the reason that the mathematical model of a scientific theory provided by the positivists is thought to clearly fit *physical* theories.

But, in fact, it fits physical theories very badly, and this for the reason that even physical theories in the usual sense – e.g. Newton's Theory of Universal Gravitation, Maxwell's theory – lead to no predictions at all without a host of auxiliary assumptions, and moreover without auxiliary assumptions that are not at all law-like, but that are, in fact, assumptions about boundary conditions and initial conditions in the case of particular systems. Thus, if the claim that the term 'gravitation', for example, had a meaning which depended on the theory were true, and the theory included such auxiliary assumptions as that 'space is a hard vacuum', and 'there is no tenth planet in the solar system', then it would follow that discovery that space is *not* a hard vacuum or even that there is a tenth planet would change the meaning of 'gravitation'. I think one has to be pretty idealistic in one's intuitions to find this at all plausible! It is not so implausible that knowledge of the meaning of the term 'gravitation' involves some knowledge of the theory (although I think that this is wrong: the stereotype associated with 'gravitation' is not nearly as strong as a particular theory of gravitation), and this is probably what most readers think of when they encounter the claim that physical magnitude terms (usually called 'theoretical terms' to prejudge just the issue this paper discusses) are 'theory loaded'; but the actual meaning-dependence required by positivist meaning theory would be a dependence not just on the *laws* of the theory, but on the particular auxiliary assumptions – for, if these are not counted as part of the theory, then the whole theory-prediction scheme collapses at the outset.

Finally, neglect of the role that auxiliary assumptions actually play in science leads to a wholly incorrect idea of how a scientific theory is confirmed. Newton's theory of gravitation was not confirmed by checking predictions derived from it plus some set of auxiliary statements fixed in advance; rather the auxiliary assumptions had to be continually modified and expanded in the history of Celestial Mechanics. That scientific problems as often have the form of finding auxiliary hypotheses as they do of finding and checking predictions is something that has been too much neglected in philosophy of science;† this neglect is largely the result of the acceptance of the positivist model and its uncritical application to actual physical theories.

† I discuss this in chapter 16, volume 1 of these papers.

12

The meaning of 'meaning'*

Language is the first broad area of human cognitive capacity for which we are beginning to obtain a description which is not exaggeratedly oversimplified. Thanks to the work of contemporary transformational linguists,† a very subtle description of at least some human languages is in the process of being constructed. Some features of these languages appear to be *universal*. Where such features turn out to be 'species-specific' – 'not explicable on some general grounds of functional utility or simplicity that would apply to arbitrary systems that serve the functions of language' – they may shed some light on the structure of mind. While it is extremely difficult to say to what extent the structure so illuminated will turn out to be a universal structure of *language*, as opposed to a universal structure of innate general learning strategies,‡ the very fact that this discussion can take place is testimony to the richness and generality of the descriptive material that linguists are beginning to provide, and also testimony to the depth of the analysis, insofar as the features that appear to be candidates for 'species-specific' features of language are in no sense surface or phenomenological features of language, but lie at the level of deep structure.

The most serious drawback to all of this analysis, as far as a philosopher is concerned, is that it does not concern the meaning of words. Analysis of the deep structure of linguistic forms gives us an incomparably more powerful description of the *syntax* of natural languages than we have ever had before. But the dimension of language associated with the word 'meaning' is, in spite of the usual spate of heroic if misguided attempts, as much in the dark as it ever was.

In this essay, I want to explore why this should be so. In my opinion, the reason that so-called semantics is in so much worse condition than syntactic theory is that the *prescientific* concept on which semantics is

* First published in K. Gunderson (ed.) *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII (University of Minnesota Press, Mpls.) © 1975 University of Minnesota.

† The contributors to this area are now too numerous to be listed: the pioneers were, of course, Zellig Harris and Noam Chomsky.

‡ For a discussion of this question see Putnam (1967) and N. Chomsky (1971), especially chapter 1.

based – the prescientific concept of *meaning* – is itself in much worse shape than the prescientific concept of syntax. As usual in philosophy, skeptical doubts about the concept do not at all help one in clarifying or improving the situation any more than dogmatic assertions by conservative philosophers that all's really well in this best of all possible worlds. The reason that the prescientific concept of meaning is in bad shape is not clarified by some general skeptical or nominalistic argument to the effect that meanings don't exist. Indeed, the upshot of our discussion will be that meanings don't exist in quite the way we tend to think they do. But electrons don't exist in quite the way Bohr thought they did, either. There is all the distance in the world between this assertion and the assertion that meanings (or electrons) 'don't exist'.

I am going to talk almost entirely about the meaning of words rather than about the meaning of sentences because I feel that our concept of word-meaning is more defective than our concept of sentence-meaning. But I will comment briefly on the arguments of philosophers such as Donald Davidson who insist that the concept of word-meaning *must* be secondary and that study of sentence-meaning must be primary. Since I regard the traditional theories about meaning as myth-eaten (notice that the topic of 'meaning' is the one topic discussed in philosophy in which there is literally nothing but 'theory' – literally nothing that can be labelled or even ridiculed as the 'common sense view'), it will be necessary for me to discuss and try to disentangle a number of topics concerning which the received view is, in my opinion, wrong. The reader will give me the greatest aid in the task of trying to make these matters clear if he will kindly assume that *nothing* is clear in advance.

Meaning and extension

Since the Middle Ages at least, writers on the theory of meaning have purported to discover an ambiguity in the ordinary concept of meaning, and have introduced a pair of terms – *extension* and *intension*, or *Sinn* and *Bedeutung*, or whatever – to disambiguate the notion. The *extension* of a term, in customary logical parlance, is simply the set of things the term is true of. Thus, 'rabbit', in its most common English sense, is true of all and only rabbits, so the extension of 'rabbit' is precisely the set of rabbits. Even this notion – and it is the *least* problematical notion in this cloudy subject – has its problems, however. Apart from problems it inherits from its parent notion of *truth*, the foregoing example of 'rabbit' *in its most common English sense* illustrates one such problem: strictly speaking, it is not a term, but an ordered pair consisting of a term and a 'sense' (or an occasion of use, or something else that dis-

tinguishes a term in one sense from the same term used in a different sense) that has an extension. Another problem is this: a 'set', in the mathematical sense, is a 'yes–no' object; any given object either definitely belongs to S or definitely does not belong to S, if S is a set. But words in a natural language are not generally 'yes–no': there are things of which the description 'tree' is clearly true and things of which the description 'tree' is clearly false, to be sure, but there are a host of borderline cases. Worse, the line between the clear cases and the borderline cases is itself fuzzy. Thus the idealization involved in the notion of *extension* – the idealization involved in supposing that there is such a thing as the set of things of which the term 'tree' is true – is actually very severe.

Recently some mathematicians have investigated the notion of a *fuzzy set* – that is, of an object to which other things belong or do not belong with a given probability or to a given degree, rather than belong 'yes–no'. If one really wanted to formalize the notion of extension as applied to terms in a natural language, it would be necessary to employ 'fuzzy sets' or something similar rather than sets in the classical sense.

The problem of a word's having more than one sense is standardly handled by treating each of the senses as a different word (or rather, by treating the word as if it carried invisible subscripts, thus: 'rabbit$_1$' – animal of a certain kind; 'rabbit$_2$' – coward; and as if 'rabbit$_1$' and 'rabbit$_2$' or whatever were different words entirely). This again involves two very severe idealizations (at least two, that is): supposing that words have discretely many senses, and supposing that the entire repertoire of senses is fixed once and for all. Paul Ziff has recently investigated the extent to which both of these suppositions distort the actual situation in natural language;† nevertheless, we will continue to make these idealizations here.

Now consider the compound terms 'creature with a heart' and 'creature with a kidney'. Assuming that every creature with a heart possesses a kidney and vice versa, the extension of these two terms is exactly the same. But they obviously differ in meaning. Supposing that there is a sense of 'meaning' in which meaning = extension, there must be another sense of 'meaning' in which the meaning of a term is not its extension but something else, say the 'concept' associated with the term. Let us call this 'something else' the *intension* of the term. The concept of a creature with a heart is clearly a different concept from the concept of a creature with a kidney. Thus the two terms have different intension. When we say they have different 'meaning', meaning = intension.

† This is discussed by Ziff (1972) especially chapter VIII.

or 'contextual'; the full form of an atomic sentence of these predicates would be '*x is five feet tall at time t*', 'x *is in pain at time* t', etc.) In science, however, it is customary to restrict the term state to properties which are defined in terms of the parameters of the individual which are fundamental from the point of view of the given science. Thus, being five feet tall is a state (from the point of view of physics); being in pain is a state (from the point of view of mentalistic psychology, at least); knowing the alphabet might be a state (from the point of view of cognitive psychology), although it is hard to say; but being a thousand miles from Paris would *not* naturally be called a *state*. In one sense, a psychological state is simply a state which is studied or described by psychology. In this sense it may be trivially true that, say *knowing the meaning of the word 'water'* is a 'psychological state' (viewed from the standpoint of cognitive psychology). But this is not the sense of psychological state that is at issue in the above assumption (I).

When traditional philosophers talked about psychological states (or 'mental' states), they made an assumption which we may call the assumption of methodological solipsism. This assumption is the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed. (In fact, the assumption was that no psychological state presupposes the existence of the subject's *body* even: if P is a psychological state, properly so called, then it must be logically possible for a 'disembodied mind' to be in P.) This assumption is pretty explicit in Descartes, but it is implicit in just about the whole of traditional philosophical psychology. Making this assumption is, of course, adopting a *restrictive program* – a program which deliberately limits the scope and nature of psychology to fit certain mentalistic preconceptions or, in some cases, to fit an idealistic reconstruction of knowledge and the world. Just *how* restrictive the program is, however, often goes unnoticed. Such common or garden variety psychological states as *being jealous* have to be reconstructed, for example, if the assumption of methodological solipsism is retained. For, in its ordinary use, *x is jealous of y* entails that *y* exists, and *x is jealous of y's regard for z* entails that both *y* and *z* exist (as well as *x*, of course). Thus *being jealous* and *being jealous of someone's regard for someone else* are not psychological states permitted by the assumption of methodological solipsism. (We shall call them 'psychological states in the wide sense' and refer to the states which are permitted by methodological solipsism as 'psychological states in the narrow sense'.) The reconstruction required by methodological solipsism would be to reconstrue *jealousy* so that I can be jealous of my own hallucinations, or of figments of my imagination, etc. Only if we assume

that psychological states in the narrow sense have a significant degree of causal closure (so that restricting ourselves to psychological states in the narrow sense will facilitate the statement of psychological *laws*) is there any point in engaging in this reconstruction, or in making the assumption of methodological solipsism. But the three centuries of failure of mentalistic psychology is tremendous evidence against this procedure, in my opinion.

Be that as it may, we can now state more precisely what we claimed at the end of the preceding section. Let A and B be any two terms which differ in extension. By assumption (II) they must differ in meaning (in the sense of 'intension'). By assumption (I), *knowing the meaning of A* and *knowing the meaning of B* are psychological states *in the narrow sense* – for this is how we shall construe assumption (I). *But these psychological states must determine the extension of the terms A and B just as much as the meanings ('intensions') do.*

To see this, let us try assuming the opposite. Of course, there cannot be two terms A and B such that *knowing the meaning of A* is the same state as *knowing the meaning of B* even though A and B have different extensions. For *knowing the meaning of A* isn't just 'grasping the intension' of A, whatever that may come to; it is also knowing that the 'intension' that one has 'grasped' *is* the intension of A. (Thus, someone who knows the meaning of 'wheel' presumably 'grasps the intension' of its German synonym *Rad*; but if he doesn't know that the 'intension' in question is the intension of *Rad* he isn't said to 'know the meaning of *Rad*'.) If A and B are different terms, then *knowing the meaning of A* is a different state from *knowing the meaning of B* whether the meanings of A and B be themselves the same or different. But by the same argument, if I_1 and I_2 are different *intensions* and A is a term, then *knowing that* I_1 *is the meaning of A* is a different psychological state from *knowing that* I_2 *is the meaning of A*. Thus, there cannot be two different logically possible worlds L_1 and L_2 such that, say, Oscar is in the *same* psychological state (in the narrow sense) in L_1 and in L_2 (in all respects), but in L_1 Oscar understands A as having the meaning I_1 and in L_2 Oscar understands A as having the meaning I_2. (For, if there were, then in L_1 Oscar would be in the psychological state *knowing that* I_1 *is the meaning of A* and in L_2 Oscar would be in the psychological state *knowing that* I_2 *is the meaning of A*, and these are different and even – assuming that A has just *one* meaning for Oscar in each world – incompatible psychological states in the narrow sense.)

In short, if S is the sort of psychological state we have been discussing – a psychological state of the form *knowing that* I *is the meaning of A*, where I is an 'intension' and A is a term – then the *same* necessary and

sufficient condition for falling into the extension of *A* 'works' in *every* logically possible world in which the speaker is in the psychological state *S*. For the state *S determines* the intension *I*, and by assumption (II) the intension amounts to a necessary and sufficient condition for membership in the *extension*.

If our interpretation of the traditional doctrine of intension and extension is fair to Frege and Carnap, then the whole psychologism/Platonism issue appears somewhat a tempest in a teapot, as far as meaning-theory is concerned. (Of course, it is a very important issue as far as general philosophy of mathematics is concerned.) For even if meanings are 'Platonic' entities rather than 'mental' entities on the Frege–Carnap view, 'grasping' those entities is presumably a psychological state (in the narrow sense). Moreover, the psychological state uniquely determines the 'Platonic' entity. So whether one takes the 'Platonic' entity or the psychological state as the 'meaning' would appear to be somewhat a matter of convention. And taking the psychological state to be the meaning would hardly have the consequence that Frege feared, that meanings would cease to be public. For psychological states are 'public' in the sense that different people (and even people in different epochs) can be in the *same* psychological state. Indeed, Frege's argument against psychologism is only an argument against identifying concepts with mental particulars, not with mental entities in general.

The 'public' character of psychological states entails, in particular, that if Oscar and Elmer understand a word *A* differently, then they must be in *different* psychological states. For the state of *knowing the intension of* A *to be, say, I* is the *same* state whether Oscar or Elmer be in it. Thus two speakers cannot be in the same psychological state in all respects and understand the term *A* differently; the psychological state of the speaker determines the intension (and hence, by assumption (II), the extension) of *A*.

It is this last consequence of the joint assumptions (I), (II) that we claim to be false. We claim that it is possible for two speakers to be in exactly the *same* psychological state (in the narrow sense), even though the extension of the term *A* in the idiolect of the one is different from the extension of the term *A* in the idiolect of the other. Extension is *not* determined by psychological state.

This will be shown in detail in later sections. If this is right, then there are two courses open to one who wants to rescue at least one of the traditional assumptions; to give up the idea that psychological state (in the narrow sense) determines *intension*, or to give up the idea that intension determines extension. We shall consider these alternatives later.

Are meanings in the head?

That psychological state does not determine extension will now be shown with the aid of a little science-fiction. For the purpose of the following science-fiction examples, we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak *English*. In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose that Twin Earth is *exactly* like Earth. He may even suppose that he has a *Doppelgänger* – an identical copy – on Twin Earth, if he wishes, although my stories will not depend on this.

Although some of the people on Twin Earth (say, the ones who call themselves 'Americans' and the ones who call themselves 'Canadians' and the ones who call themselves 'Englishmen', etc.) speak English, there are, not surprisingly, a few tiny differences which we will now describe between the dialects of English spoken on Twin Earth and Standard English. These differences themselves depend on some of the peculiarities of Twin Earth.

One of the peculiarities of Twin Earth is that the liquid called 'water' is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as *XYZ*. I shall suppose that *XYZ* is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain *XYZ* and not water, that it rains *XYZ* on Twin Earth and not water, etc.

If a spaceship from Earth ever visits Twin Earth, then the supposition at first will be that 'water' has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that 'water' on Twin Earth is *XYZ*, and the Earthian spaceship will report somewhat as follows:

'On Twin Earth the word "water" means *XYZ*.'

(It is this sort of use of the word 'means' which accounts for the doctrine that extension is one sense of 'meaning', by the way. But note that although 'means' does mean something like *has as extension* in this example, one would *not* say

'On Twin Earth the meaning of the word "water" is *XYZ*.'

unless, possibly, the fact that 'water is *XYZ*' was known to every adult speaker of English on Twin Earth. We can account for this in terms of the theory of meaning we develop below; for the moment we just

remark that although the verb 'means' sometimes means 'has as extension', the nominalization 'meaning' *never* means 'extension'.)

Symmetrically, if a spaceship from Twin Earth ever visits Earth, then the supposition at first will be that the word 'water' has the same meaning on Twin Earth and on Earth. This supposition will be corrected when it is discovered that 'water' on Earth is H_2O, and the Twin Earthian spaceship will report:

'On Earth† the word "water" means H_2O.'

Note that there is no problem about the extension of the term 'water'. The word simply has two different meanings (as we say): in the sense in which it is used on Twin Earth, the sense of $water_{TE}$, what *we* call 'water' simply isn't water; while in the sense in which it is used on Earth, the sense of $water_E$, what the Twin Earthians call 'water' simply isn't water. The extension of 'water' in the sense of $water_E$ is the set of all wholes consisting of H_2O molecules, or something like that; the extension of water in the sense of $water_{TE}$ is the set of all wholes consisting of XYZ molecules, or something like that.

Now let us roll the time back to about 1750. At that time chemistry was not developed on either Earth or Twin Earth. The typical Earthian speaker of English did not know water consisted of hydrogen and oxygen, and the typical Twin Earthian speaker of English did not know 'water' consisted of XYZ. Let $Oscar_1$ be such a typical Earthian English speaker, and let $Oscar_2$ be his counterpart on Twin Earth. You may suppose that there is no belief that $Oscar_1$ had about water that $Oscar_2$ did not have about 'water'. If you like, you may even suppose that $Oscar_1$ and $Oscar_2$ were exact duplicates in appearance, feelings, thoughts, interior monologue, etc. Yet the extension of the term 'water' was just as much H_2O on Earth in 1750 as in 1950; and the extension of the term 'water' was just as much XYZ on Twin Earth in 1750 as in 1950. $Oscar_1$ and $Oscar_2$ understood the term 'water' differently in 1750 *although they were in the same psychological state*, and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood the term 'water' differently. Thus the extension of the term 'water' (and, in fact, its 'meaning' in the intuitive preanalytical usage of that term) is *not* a function of the psychological state of the speaker by itself.

But, it might be objected, why should we accept it that the term 'water' has the same extension in 1750 and in 1950 (on both Earths)? The logic of natural-kind terms like 'water' is a complicated matter,

† Or rather, they will report: 'On Twin Earth (*the Twin Earthian name for Terra* – H.P.) the word "water" means H_2O.'

but the following is a sketch of an answer. Suppose I point to a glass of water and say 'this liquid is called water' (or 'this is called water', if the marker 'liquid' is clear from the context). My 'ostensive definition' of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation (say, *x is the same liquid as y*, or *x is the* $same_L$ *as y*) to most of the stuff I and other speakers in my linguistic community have on other occasions called 'water'. If this presupposition is false because, say, I am without knowing it pointing to a glass of gin and not a glass of water, then I do not intend my ostensive definition to be accepted. Thus the ostensive definition conveys what might be called a defeasible necessary and sufficient condition: the necessary and sufficient condition for being water is bearing the relation $same_L$ to the stuff in the glass; but this is the necessary and sufficient condition only if the empirical presupposition is satisfied. If it is not satisfied, then one of a series of, so to speak, 'fallback' conditions becomes activated.

The key point is that the relation $same_L$ is a *theoretical* relation: whether something is or is not the same liquid as *this* may take an indeterminate amount of scientific investigation to determine. Moreover, even if a 'definite' answer has been obtained either through scientific investigation or through the application of some 'common sense' test, the answer is *defeasible*: future investigation might reverse even the most 'certain' example. Thus, the fact that an English speaker in 1750 might have called XYZ 'water', while he or his successors would not have called XYZ water in 1800 or 1850 does not mean that the 'meaning' of 'water' changed for the average speaker in the interval. In 1750 or in 1850 or in 1950 one might have pointed to, say, the liquid in Lake Michigan as an example of 'water'. What changed was that in 1750 we would have mistakenly thought that XYZ bore the relation $same_L$ to the liquid in Lake Michigan, while in 1800 or 1850 we would have known that it did not (I am ignoring the fact that the liquid in Lake Michigan was only dubiously water in 1950, of course).

Let us now modify our science-fiction story. I do not know whether one can make pots and pans out of molybdenum; and if one can make them out of molybdenum, I don't know whether they could be distinguished easily from aluminum pots and pans. (I don't know any of this even though I have acquired the word 'molybdenum'.) So I shall suppose that molybdenum pots and pans *can't* be distinguished from aluminum pots and pans save by an expert. (To emphasize the point, I repeat that this could be true for all I know, and *a fortiori* it could be true for all I know by virtue of 'knowing the meaning' of the words *aluminum* and *molybdenum*.) We will now suppose that molybdenum is

as common on Twin Earth as aluminum is on Earth, and that aluminum is as rare on Twin Earth as molybdenum is on Earth. In particular, we shall assume that 'aluminum' pots and pans are made of molybdenum on Twin Earth. Finally, we shall assume that the words 'aluminum' and 'molybdenum' are *switched* on Twin Earth: 'aluminum' is the name of *molybdenum* and 'molybdenum' is the name of *aluminum*.

This example shares some features with the previous one. If a spaceship from Earth visited Twin Earth, the visitors from Earth probably would not suspect that the 'aluminum' pots and pans on Twin Earth were not made of aluminum, especially when the Twin Earthians *said* they were. But there is one important difference between the two cases. An Earthian metallurgist could tell very easily that 'aluminum' was molybdenum, and a Twin Earthian metallurgist could tell equally easily that aluminum was 'molybdenum'. (The shudder quotes in the preceding sentence indicate Twin Earthian usages.) Whereas in 1750 no one on either Earth or Twin Earth could have distinguished water from 'water', the confusion of aluminum with 'aluminum' involves only a part of the linguistic communities involved.

The example makes the same point as the preceding one. If Oscar_1 and Oscar_2 are standard speakers of Earthian English and Twin Earthian English respectively, and neither is chemically or metallurgically sophisticated, then there may be no difference at all in their psychological state when they use the word 'aluminum'; nevertheless we have to say that 'aluminum' has the extension *aluminum* in the idiolect of Oscar_1 and the extension *molybdenum* in the idiolect of Oscar_2. (Also we have to say that Oscar_1 and Oscar_2 mean different things by 'aluminum', that 'aluminum' has a different meaning on Earth than it does on Twin Earth, etc.) Again we see that the psychological state of the speaker does *not* determine the extension (*or* the 'meaning', speaking preanalytically) of the word.

Before discussing this example further, let me introduce a *non*-science-fiction example. Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of 'beech' in *both* of our idiolects. Thus 'elm' in my idiolect has a different extension from 'beech' in your idiolect (as it should). Is it really credible that this difference in extension is brought about by some difference in our *concepts*? My *concept* of an elm tree is exactly the same as my concept of a beech tree (I blush to confess). (This shows that the identification of meaning 'in the sense of intension' with *concept* cannot be correct, by the way.) If someone heroically attempts to maintain that the difference

between the extension of 'elm' and the extension of 'beech' in *my* idiolect is explained by a difference in my psychological state, then we can always refute him by constructing a 'Twin Earth' example – just let the words 'elm' and 'beech' be switched on Twin Earth (the way 'aluminum' and 'molybdenum' were in the previous example). Moreover, I suppose I have a *Doppelgänger* on Twin Earth who is molecule for molecule 'identical' with me (in the sense in which two neckties can be 'identical'). If you are a dualist, then also suppose my *Doppelgänger* thinks the same verbalized thoughts I do, has the same sense data, the same dispositions, etc. It is absurd to think *his* psychological state is one bit different from mine: yet he 'means' *beech* when he says 'elm' and *I* 'mean' *elm* when I say elm. Cut the pie any way you like, 'meanings' just ain't in the *head*!

A socio-linguistic hypothesis

The last two examples depend upon a fact about language that seems, surprisingly, never to have been pointed out: that there is *division of linguistic labor*. We could hardly use such words as 'elm' and 'aluminum' if no one possessed a way of recognizing elm trees and aluminum metal; but not everyone to whom the distinction is important has to be able to make the distinction. Let us shift the example: consider *gold*. Gold is important for many reasons: it is a precious metal, it is a monetary metal, it has symbolic value (it is important to most people that the 'gold' wedding ring they wear *really* consist of gold and not just *look* gold), etc. Consider our community as a 'factory': in this 'factory' some people have the 'job' of *wearing gold wedding rings*, other people have the 'job' of *selling gold wedding rings*, still other people have the 'job' of *telling whether or not something is really gold*. It is not at all necessary or efficient that everyone who wears a gold ring (or a gold cufflink, etc.), or discusses the 'gold standard', etc., engage in buying and selling gold. Nor is it necessary or efficient that everyone who buys and sells gold be able to tell whether or not something is really gold in a society where this form of dishonesty is uncommon (selling fake gold) and in which one can easily consult an expert in case of doubt. And it is *certainly* not necessary or efficient that everyone who has occasion to buy or wear gold be able to tell with any reliability whether or not something is really gold.

The foregoing facts are just examples of mundane division of labor (in a wide sense). But they engender a division of linguistic labor: everyone to whom gold is important for any reason has to *acquire* the word 'gold'; but he does not have to acquire the *method of recognizing*

if something is or is not gold. He can rely on a special subclass of speakers. The features that are generally thought to be present in connection with a general name – necessary and sufficient conditions for membership in the extension, ways of recognizing if something is in the extension ('criteria'), etc. – are all present in the linguistic community *considered as a collective body;* but that collective body divides the 'labor' of knowing and employing these various parts of the 'meaning' of 'gold'.

This division of linguistic labor rests upon and presupposes the division of *non*linguistic labor, of course. If only the people who know how to tell if some metal is really gold or not have any reason to have the word 'gold' in their vocabulary, then the word 'gold' will be as the word 'water' was in 1750 with respect to that subclass of speakers, and the other speakers just won't acquire it at all. And some words do not exhibit any division of linguistic labor: 'chair', for example. But with the increase of division of labor in the society and the rise of science, more and more words begin to exhibit this kind of division of labor. 'Water', for example, did not exhibit it at all prior to the rise of chemistry. Today it is obviously necessary for every speaker to be able to recognize water (reliably under normal conditions), and probably every adult speaker even knows the necessary and sufficient condition 'water is H_2O', but only a few adult speakers could distinguish water from liquids which superficially resembled water. In case of doubt, other speakers would rely on the judgement of these 'expert' speakers. Thus the way of recognizing possessed by these 'expert' speakers is also, through them, possessed by the collective linguistic body, even though it is not possessed by each individual member of the body, and in this way the most recherché fact about water may become part of the *social* meaning of the word while being unknown to almost all speakers who acquire the word.

It seems to me that this phenomenon of division of linguistic labor is one which it will be very important for sociolinguistics to investigate. In connection with it, I should like to propose the following hypothesis:

HYPOTHESIS OF THE UNIVERSALITY OF THE DIVISION OF LINGUISTIC LABOR: Every linguistic community exemplifies the sort of division of linguistic labor just described: that is, possesses at least some terms whose associated 'criteria' are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets.

It would be of interest, in particular, to discover if extremely primitive peoples were sometimes exceptions to this hypothesis (which would indicate that the division of linguistic labor is a product of social evolution), or if even they exhibit it. In the latter case, one might conjecture that division of labor, including linguistic labor, is a fundamental trait of our species.

It is easy to see how this phenomenon accounts for some of the examples given above of the failure of the assumptions (1), (2). Whenever a term is subject to the division of linguistic labor, the 'average' speaker who acquires it does not acquire anything that fixes its extension. In particular, his individual psychological state *certainly* does not fix its extension; it is only the sociolinguistic state of the collective linguistic body to which the speaker belongs that fixes the extension.

We may summarize this discussion by pointing out that there are two sorts of tools in the world: there are tools like a hammer or a screwdriver which can be used by one person; and there are tools like a steamship which require the cooperative activity of a number of persons to use. Words have been thought of too much on the model of the first sort of tool.

Indexicality and rigidity†

The first of our science-fiction examples – 'water' on Earth and on Twin Earth in 1750 – does not involve division of linguistic labor, or at least does not involve it in the same way the examples of 'aluminum' and 'elm' do. There were not (in our story, anyway) any 'experts' on water on Earth in 1750, nor any experts on 'water' on Twin Earth. (The example *can* be construed as involving division of labor *across time*, however. I shall not develop this method of treating the example here.) The example *does* involve things which are of fundamental importance to the theory of reference and also to the theory of necessary truth, which we shall now discuss.

There are two obvious ways of telling someone what one means by a natural-kind term such as 'water' or 'tiger' or 'lemon'. One can give him a so-called ostensive definition – 'this (liquid) is water'; 'this (animal) is a tiger'; 'this (fruit) is a lemon'; where the parentheses are meant to indicate that the 'markers' *liquid, animal, fruit,* may be either explicit or implicit. Or one can give him a *description*. In the latter case the description one gives typically consists of one or more markers

† The substance of this section was presented at a series of lectures I gave at the University of Washington (Summer Institute in Philosophy) in 1968, and at a lecture at the University of Minnesota.

together with a *stereotype* (see chapter 8 in this volume) – a standardized description of features of the kind that are typical, or 'normal', or at any rate stereotypical. The central features of the stereotype generally are *criteria* – features which in normal situations constitute ways of recognizing if a thing belongs to the kind or, at least, necessary conditions (or probabilistic necessary conditions) for membership in the kind. Not all criteria used by the linguistic community as a collective body are included in the stereotype, and in some cases the stereotypes may be quite weak. Thus (unless I am a very atypical speaker), the stereotype of an elm is just that of a common deciduous tree. These features are indeed necessary conditions for membership in the kind (I mean 'necessary' in a loose sense; I don't think 'elm trees are deciduous' is *analytic*), but they fall far short of constituting a way of recognizing elms. On the other hand, the stereotype of a tiger does enable one to recognize tigers (unless they are albino, or some other atypical circumstance is present), and the stereotype of a lemon generally enables one to recognize lemons. In the extreme case, the stereotype may be *just* the marker: the stereotype of molybdenum might be *just* that molybdenum is a *metal*. Let us consider both of these ways of introducing a term into someone's vocabulary.

Suppose I point to a glass of liquid and say '*this* is water', in order to teach someone the word 'water'. We have already described some of the empirical presuppositions of this act, and the way in which this kind of meaning-explanation is defeasible. Let us now try to clarify further how it is supposed to be taken.

In what follows, we shall take the notion of 'possible world' as primitive. We do this because we feel that in several senses the notion makes sense and is scientifically important even if it needs to be made more precise. We shall assume further that in at least some cases it is possible to speak of the same individual as existing in more than one possible world.† Our discussion leans heavily on the work of Saul Kripke, although the conclusions were obtained independently.

Let W_1 and W_2 be two possible worlds in which I exist and in which this glass exists and in which I am giving a meaning explanation by pointing to this glass and saying 'this is water'. (We do *not* assume that the *liquid* in the glass is the same in both worlds.) Let us suppose that in W_1 the glass is full of H_2O and in W_2 the glass is full of XYZ. We shall also suppose that W_1 is the *actual* world and that XYZ is the stuff typically called 'water' in the world W_2 (so that the relation between English speakers in W_1 and English speakers in W_2 is exactly the same

† This assumption is not actually needed in what follows. What *is* needed is that the same *natural kind* can exist in more than one possible world.

as the relation between English speakers on Earth and English speakers on Twin Earth). Then there are two theories one might have concerning the meaning of 'water'.

(1) One might hold that 'water' was *world-relative* but *constant* in meaning (i.e. the word has a *constant relative meaning*). In this theory, 'water' *means the same* in W_1 and W_2; it's just that water is H_2O in W_1 and water is XYZ in W_2.

(2) One might hold that water is H_2O in all worlds (the stuff called 'water' in W_2 isn't water), but 'water' doesn't have the same meaning in W_1 and W_2.

If what was said before about the Twin Earth case was correct, then (2) is clearly the correct theory. When I say '*this* (liquid) is water', the 'this' is, so to speak, a *de re* 'this' – i.e. the force of my explanation is that 'water' is whatever bears a certain equivalence relation (the relation we called 'same$_L$' above) to the piece of liquid referred to as 'this' *in the actual world.*

We might symbolize the difference between the two theories as a 'scope' difference in the following way. In theory (1), the following is true:

(1′) (For every world W) (For every x in W) (x is water $\equiv$ x bears same$_L$ to the entity referred to as 'this' in W)

while on theory (2):

(2′) (For every world W) (For every x in W) (x is water $\equiv$ x bears same$_L$ to the entity referred to as 'this' *in the actual world* W_1).

(I call this a 'scope' difference because in (1′) 'the entity referred to as "this"' is within the scope of 'For every world W' – as the qualifying phrase 'in W' makes explicit, whereas in (2′) 'the entity referred to as "this"' means 'the entity referred to as "this" *in the actual world*', and has thus a reference *independent* of the bound variable 'W'.)

Kripke calls a designator 'rigid' (in a given sentence) if (in that sentence) it refers to the same individual in every possible world in which the designator designates. If we extend the notion of rigidity to substance names, then we may express Kripke's theory and mine by saying that the term 'water' is *rigid.*

The rigidity of the term 'water' follows from the fact that when I give the ostensive definition '*this* (liquid) is water' I intend (2′) and not (1′).

We may also say, following Kripke, that when I give the ostensive definition '*this* (liquid) is water', the demonstrative 'this' is *rigid.*

What Kripke was the first to observe is that this theory of the meaning (or 'use', or whatever) of the word 'water' (and other natural-kind terms as well) has startling consequences for the theory of necessary truth.

To explain this, let me introduce the notion of a *cross-world relation*. A two-term relation R will be called *cross-world* when it is understood in such a way that its extension is a set of ordered pairs of individuals *not all in the same possible world*. For example, it is easy to understand the relation *same height as* as a cross-world relation: just understand it so that, e.g. if x is an individual in a world W_1 who is five feet tall (in W_1) and y is an individual in W_2 who is five feet tall (in W_2), then the ordered pair x, y belongs to the extension of *same height as*. (Since an individual may have different heights in different possible worlds in which that same individual exists, strictly speaking it is not the ordered pair x, y that constitutes an element of the extension of *same height as*, but rather the ordered pair *x-in-world-*W_1, *y-in-world-*W_2.)

Similarly, we can understand the relation $same_L$ (same liquid as) as a cross-world relation by understanding it so that a liquid in world W_1 which has the same important physical properties (in W_1) that a liquid in W_2 possesses (in W_2) bears $same_L$ to the latter liquid.

Then the theory we have been presenting may be summarized by saying that an entity x, in an arbitrary possible world, is *water* if and only if it bears the relation $same_L$ (construed as a cross-world relation) to the stuff *we* call 'water' in the *actual* world.

Suppose, now, that I have not yet discovered what the important physical properties of water are (in the actual world) – i.e. I don't yet know that water is H_2O. I may have ways of *recognizing* water that are successful (of course, I may make a small number of mistakes that I won't be able to detect until a later stage in our scientific development) but not know the microstructure of water. If I agree that a liquid with the superficial properties of 'water' but a different microstructure *isn't really water*, then my ways of recognizing water (my 'operational definition', so to speak) cannot be regarded as an analytical specification of *what it is to be* water. Rather, the operational definition, like the ostensive one, is simply a way of pointing out a standard – pointing out the stuff *in the actual world* such that for x to be water, in *any* world, is for x to bear the relation $same_L$ to the *normal* members of the class of *local* entities that satisfy the operational definition. 'Water' on Twin Earth is not water, even if it satisfies the operational definition, because it doesn't bear $same_L$ to the *local* stuff that satisfies the operational definition, and local stuff that satisfies the operational definition but has a microstructure different from rest of the local stuff that satisfies the

operational definition isn't water either, because it doesn't bear $same_L$ to the *normal* examples of the local 'water'.

Suppose, now, that I discover the microstructure of water – that water is H_2O. At this point I will be able to say that the stuff on Twin Earth that I earlier *mistook* for water isn't really water. In the same way, if you describe not another planet in the actual universe, but another possible universe in which there is stuff with the chemical formula XYZ which passes the 'operational test' for *water*, we shall have to say that that stuff isn't water but merely XYZ. You will not have described a possible world in which 'water is XYZ', but merely a possible world in which there are lakes of XYZ, people drink XYZ (and not water), or whatever. In fact, once we have discovered the nature of water, nothing counts as a possible world in which water doesn't have that nature. Once we have discovered that water (in the actual world) is H_2O, *nothing counts as a possible world in which water isn't* H_2O. In particular, if a 'logically possible' statement is one that holds in some 'logically possible world', *it isn't logically possible that water isn't* H_2O.

On the other hand, we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H_2O. In that sense, it is conceivable that water isn't H_2O. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.

Kripke refers to statements which are rationally unrevisable (assuming there are such) as *epistemically necessary*. Statements which are true in all possible worlds he refers to simply as necessary (or sometimes as 'metaphysically necessary'). In this terminology, the point just made can be restated as: a statement can be (metaphysically) necessary and epistemically contingent. Human intuition has no privileged access to metaphysical necessity.

Since Kant there has been a big split between philosophers who thought that all necessary truths were analytic and philosophers who thought that some necessary truths were synthetic *a priori*. But none of these philosophers thought that a (metaphysically) necessary truth could fail to be *a priori*: the Kantian tradition was as guilty as the empiricist tradition of equating metaphysical and epistemic necessity. In this sense Kripke's challenge to received doctrine goes far beyond the usual empiricism/Kantianism oscillation.

In this paper our interest is in theory of meaning, however, and not in theory of necessary truth. Points closely related to Kripke's have been made in terms of the notion of *indexicality*.† Words like 'now', 'this',

† These points were made in my 1968 lectures at the University of Washington and the University of Minnesota.

'here', have long been recognized to be *indexical*, or *token-reflexive* – i.e. to have an extension which varied from context to context or token to token. For these words no one has ever suggested the traditional theory that 'intension determines extension'. To take our Twin Earth example: if I have a *Doppelgänger* on Twin Earth, then when I think 'I have a headache', *he* thinks 'I have a headache'. But the extension of the particular token of 'I' in his verbalized thought is himself (or his unit class, to be precise), while the extension of the token of 'I' in *my* verbalized thought is *me* (or my unit class, to be precise). So the same word, 'I', has two different extensions in two different idiolects; but it does not follow that the concept I have of myself is in any way different from the concept my *Doppelgänger* has of himself.

Now then, we have maintained that indexicality extends beyond the *obviously* indexical words and morphemes (e.g. the tenses of verbs). Our theory can be summarized as saying that words like 'water' have an unnoticed indexical component: 'water' is stuff that bears a certain similarity relation to the water *around here*. Water at another time or in another place or even in another possible world has to bear the relation same_L to *our* 'water' *in order to be water*. Thus the theory that (1) words have 'intensions', which are something like concepts associated with the words by speakers; and that (2) intension determines extension – cannot be true of natural-kind words like 'water' for the same reason the theory cannot be true of obviously indexical words like 'I'.

The theory that natural-kind words like 'water' are indexical leaves it open, however, whether to say that 'water' in the Twin Earth dialect of English has the same *meaning* as 'water' in the Earth dialect and a different extension (which is what we normally say about 'I' in different idiolects), thereby giving up the doctrine that 'meaning (intension) determines extension'; or to say, as we have chosen to do, that difference is extension is *ipso facto* a difference in meaning for natural-kind words, thereby giving up the doctrine that meanings are concepts, or, indeed, mental entities of *any* kind.

It should be clear, however, that Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point. We heartily endorse what Kripke says when he writes:

> Let us suppose that we do fix the reference of a name by a description. Even if we do so, we do not then make the name synonymous with the description, but instead we use the name rigidly to refer to the object so named, even in talking about counterfactual situations where the thing named would not satisfy the description in question. Now, this is what I think is in fact true for

> those cases of naming where the reference is fixed by description. But, in fact, I also think, contrary to most recent theorists, that the reference of names is rarely or almost never fixed by means of description. And by this I do not just mean what Searle says: 'It's not a single description, but rather a cluster, a family of properties that fixes the reference.' I mean that properties in this sense are not used at all. (Kripke, 1972, p. 157)

Let's be realistic

I wish now to contrast my view with one which is popular, at least among students (it appears to arise spontaneously). For this discussion, let us take as our example of a natural-kind word the word *gold*. We will not distinguish between 'gold' and the cognate words in Greek, Latin etc. And we will focus on 'gold' in the sense of gold in the solid state. With this understood, we maintain: 'gold' has not changed its *extension* (or not changed it significantly) in two thousand years. Our methods of *identifying* gold have grown incredibly sophisticated. But the extension of χρυσὸς in Archimedes' dialect of Greek is the same as the extension of *gold* in my dialect of English.

It is possible (and let us suppose it to be the case) that just as there were pieces of metal which could not have been determined *not* to be gold prior to Archimedes, so there were or are pieces of metal which could not have been determined *not* to be gold in Archimedes' day, but which we can distinguish from gold quite easily with modern techniques. Let X be such a piece of metal. Clearly X does not lie in the extension of 'gold' in standard English; my view is that it did not lie in the extension of χρυσὸς in Attic Greek, either, although an ancient Greek would have *mistaken* X for gold (or, rather, χρυσὸς).

The alternative view is that 'gold' *means* whatever satisfies the *contemporary* 'operational definition' of *gold*. 'Gold' a hundred years ago meant whatever satisfied the 'operational definition' of *gold* in use a hundred years ago; 'gold' now means whatever satisfies the operational definition of *gold* in use in 1973; and χρυσὸς meant whatever satisfied the operational definition of χρυσὸς in use *then*.

One common motive for adopting this point of view is a certain skepticism about *truth*. In the view I am advocating, when Archimedes asserted that something was gold (χρυσὸς) he was not just saying that it had the superficial characteristics of gold (in exceptional cases, something may belong to a natural kind and *not* have the superficial characteristics of a member of that natural kind, in fact); he was saying that it had the same general *hidden structure* (the same 'essence', so to speak) as any normal piece of local gold. Archimedes would have said

that our hypothetical piece of metal X was gold, but he would have been *wrong*. But *who's to say* he would have been wrong?

The obvious answer is: *we are* (using the best theory available today). For most people either the question (*who's to say?*) has bite, and our answer has no bite, or our answer has bite and the question has no bite. Why is this?

The reason, I believe, is that people tend either to be strongly anti-realistic or strongly realistic in their intuitions. To a strongly anti-realistic intuition it makes little sense to say that what is in the extension of Archimedes' term χρυσὸς is to be determined using *our* theory. For the antirealist does not see our theory and Archimedes' theory as two approximately correct descriptions of some fixed realm of theory-independent entities, and he tends to be skeptical about the idea of 'convergence' in science – he does not think our theory is a *better* description of the *same* entities that Archimedes was describing. But if our theory is *just* our theory, then to use *it* in deciding whether or not X lies in the extension of χρυσὸς is just as arbitrary as using Neanderthal theory to decide whether or not X lies in the extension of χρυσὸς. The only theory that it is *not* arbitrary to use is the one the speaker himself subscribes to.

The trouble is that for a strong antirealist *truth* makes no sense except as an intra-theoretic notion (see the preceding chapter for a discussion of this point). The antirealist can use truth intra-theoretically in the sense of a 'redundancy theory'; but he does not have the notions of truth and reference available *extra-theoretically*. But *extension is tied to the notion of truth*. The extension of a term is just what the term is *true of*. Rather than try to retain the notion of extension via an awkward operationalism, the antirealist should reject the notion of extension as he does the notion of truth (in any extra-theoretic sense). Like Dewey, for example, he can fall back on a notion of 'warranted assertibility' instead of truth (relativized to the scientific method, if he thinks there is a *fixed* scientific method, or to the best methods available at the time, if he agrees with Dewey that the scientific method itself evolves). Then he can say that 'X is gold (χρυσὸς)' was warrantedly assertible in Archimedes' time and is not warrantedly assertible today (indeed, this is a *minimal* claim, in the sense that it represents the minimum that the realist and the antirealist can agree on); but the assertion that X was in the extension of χρυσὸς will be rejected as meaningless, like the assertion that 'X is gold (χρυσὸς)' was *true*.

It is well known that narrow operationalism cannot successfully account for the actual use of scientific or common-sense terms. Loosened versions of operationalism, like Carnap's version of Ramsey's theory,

agree with, if they do not account for, actual scientific use (mainly because the loosened versions agree with any possible use!), but at the expense of making the communicability of scientific results a *miracle*. It is beyond question that scientists use terms as if the associated criteria were not *necessary and sufficient conditions*, but rather *approximately* correct characterizations of some world of theory-independent entities, and that they talk as if later theories in a mature science were, in general, *better* descriptions of the *same* entities that earlier theories referred to. In my opinion the hypothesis that this is *right* is the only hypothesis that can account for the communicability of scientific results, the closure of acceptable scientific theories under first-order logic, and many other features of the scientific method.† But it is not my task to argue this here. My point is that if we are to use the notions of truth and extension in an extra-theoretic way (i.e. to regard those notions as defined for statements couched in the languages of theories other than our own), then we should accept the realist perspective to which those notions belong. The doubt about whether *we* can say that X does not lie in the extension of 'gold' as *Jones* used it is the *same* doubt as the doubt whether it makes sense to think of Jones's statement that 'X is gold' as *true or false* (and not just 'warrantedly assertible for Jones and not warrantedly assertible for us'). To square the notion of truth, which is essentially a realist notion, with one's antirealist prejudices by adopting an untenable theory of meaning is no progress.

A second motive for adopting an extreme operationalist account is a dislike of unverifiable hypotheses. At first blush it may seem as if we are saying that 'X is gold (χρυσὸς)' was false in Archimedes' time although Archimedes could not *in principle* have known that it was false. But this is not exactly the situation. The fact is that there are a host of situations that *we* can describe (using the very theory that tells us that X isn't gold) in which X would have behaved quite unlike the rest of the stuff Archimedes classified as gold. Perhaps X would have separated into two different metals when melted, or would have had different conductivity properties, or would have vaporized at a different temperature, or whatever. If we had performed the experiments with Archimedes watching, he might not have known the theory, but he would have been able to check the empirical regularity that 'X behaves differently from the rest of the stuff I classify as χρυσὸς in several respects'. Eventually he would have concluded that 'X may not be gold'.

The point is that even if something satisfies the criteria used at a

† For an illuminating discussion of just these points, see R. Boyd's *Realism and Scientific Epistemology* (unpublished: Xerox draft circulated by author, Cornell Dept. of Philosophy).

given time to identify gold (i.e., to recognize if something is gold), it may behave differently in one or more situations from the rest of the stuff that satisfies the criteria. This may not *prove* that it isn't gold, but it puts the hypothesis that it may not be gold in the running , even in the absence of theory. If, now, we had gone on to inform Archimedes that gold had such and such a molecular structure (except for X), and that X behaved differently because it had a different molecular structure, is there any doubt that he would have agreed with us that X isn't gold? In any case, to worry because things may be *true* (at a given time) that can't be *verified* (at that time) seems to me ridiculous. In any reasonable view there are surely things that are true and can't be verified at *any* time. For example, suppose there are infinitely many binary stars. *Must* we be able to verify this, even *in principle*? (See chapter 22 in this volume, and chapters 17 and 18, volume 1.)

So far we have dealt with *metaphysical* reasons for rejecting our account. But someone might disagree with us about the empirical facts concerning the intentions of speakers. This would be the case if, for instance, someone thought that Archimedes (in the *Gedankenexperiment* described above) would have said: 'it doesn't matter if X *does* act differently from other pieces of gold; X is a piece of gold, because X has such-and-such properties and that's all it takes to be gold'. While, indeed, we cannot be certain that natural-kind words in ancient Greek had the properties of the corresponding words in present-day English, there cannot be any serious doubt concerning the properties of the latter. If we put philosophical prejudices aside, then I believe that we know perfectly well that no operational definition does provide a necessary and sufficient condition for the application of any such word. We may give an 'operational definition', or a cluster of properties, or whatever, but the intention is never to 'make the name *synonymous* with the description'. Rather 'we use the name *rigidly*' to refer to whatever things share the *nature* that things satisfying the description normally possess.

Other senses

What we have analyzed so far is the predominant sense of natural-kind words (or, rather, the predominant *extension*). But natural-kind words typically possess a number of senses. (Ziff has even suggested that they possess a *continuum* of senses.)

Part of this can be explained on the basis of our theory. To be water, for example, is to bear the relation same_L to certain things. But what is the relation same_L?

x bears the relation same_L to y just in case (1) x and y are both liquids,

and (2) x and y agree in important physical properties. The term 'liquid' is itself a natural-kind term that I shall not try to analyze here. The term 'property' is a broad-spectrum term that we have analyzed in previous papers. What I want to focus on now is the notion of *importance*. Importance is an interest-relative notion. Normally the 'important' properties of a liquid or solid, etc., are the ones that are *structurally* important: the ones that specify what the liquid or solid, etc., is ultimately made out of – elementary particles, or hydrogen and oxygen, or earth, air, fire, water, or whatever – and how they are arranged or combined to produce the superficial characteristics. From this point of view the characteristic of a typical bit of water is consisting of H_2O. But it may or may not be important that there are impurities; thus, in one context 'water' may mean *chemically pure water*, while in another it may mean the stuff in Lake Michigan. And a speaker may sometimes refer to XYZ as water if one is *using* it as water. Again, normally it is important that water is in the liquid state; but sometimes it is unimportant, and one may refer to a single H_2O molecule as water, or to water vapor as water ('water in the air').

Even senses that are so far out that they have to be regarded as a bit 'deviant' may bear a definite relation to the core sense. For example, I might say 'did you see the lemon', meaning the *plastic* lemon. A less deviant case is this: we discover 'tigers' on Mars. That is, they look just like tigers, but they have a silicon-based chemistry instead of a carbon-based chemistry. (A remarkable example of parallel evolution!) Are Martian 'tigers' tigers? It depends on the context.

In the case of this theory, as in the case of any theory that is orthogonal to the way people have thought about something previously, misunderstandings are certain to arise. One which has already arisen is the following: a critic has maintained that the *predominant* sense of, say, 'lemon' is the one in which anything with (a sufficient number of) the superficial characteristics of a lemon is a lemon. The same critic has suggested that having the hidden structure – the genetic code – of a lemon is necessary to being a lemon only when 'lemon' is used as a term of *science*. Both of these contentions seem to me to rest on a misunderstanding, or, perhaps, a pair of complementary misunderstandings.

The sense in which literally *anything* with the superficial characteristics of a lemon is necessarily a lemon, far from being the dominant one, is extremely deviant. In that sense something would be a lemon if it looked and tasted like a lemon, even if it had a silicon-based chemistry, for example, or even if an electron-microscope revealed it to be a *machine*. (Even if we include growing 'like a lemon' in the superficial

characteristics, this does not exclude the silicon lemon, if there are 'lemon' trees on Mars. It doesn't even exclude the machine-lemon; maybe the tree is a machine too!)

At the same time the sense in which to be a lemon something has to have the genetic code of a lemon is *not* the same as the technical sense (if there is one, which I doubt). The technical sense, I take it, would be one in which 'lemon' was *synonymous* with a description which *specified* the genetic code. But when we said (to change the example) that to be *water* something has to be H_2O we did not mean, as we made clear, that the *speaker* has to *know* this. It is only by confusing *metaphysical* necessity with *epistemological* necessity that one can conclude that, if the (metaphysically necessary) truth-condition for being water is being H_2O, then 'water' must be synonymous with H_2O – in which case it is certainly a term of science. And similarly, even though the predominant sense of 'lemon' is one in which to be a lemon something has to have the genetic code of a lemon (I believe), it does not follow that 'lemon' is synonymous with a description which specifies the genetic code explicitly or otherwise.

The mistake of thinking that there is an important sense of 'lemon' (perhaps the predominant one) in which to have the superficial characteristics of a lemon is at least *sufficient* for being a lemon is more plausible if among the superficial characteristics one includes *being cross-fertile with lemons*. But the characteristic of being cross-fertile with lemons presupposes the notion of being a lemon. Thus, even if one can obtain a sufficient condition in *this* way, to take this as inconsistent with the characterization offered here is question-begging. Moreover the characterization in terms of *lemon*-presupposing 'superficial characteristics' (like being cross-fertile with *lemons*) gives no truth-condition which would enable us to decide which objects in other possible worlds (or which objects a million years ago, or which objects a million light years from here) are lemons. (In addition, I don't think this characterization, question-begging as it is, is *correct*, even as a sufficient condition. I think one could invent cases in which something which was not a lemon was cross-fertile with lemons and looked like a lemon, etc.)

Again, one might try to rule out the case of the machine-lemon (lemon-machine?) which 'grows' on a machine-tree (tree-machine?) by saying that 'growing' is not really *growing*. That is right; but it's right because *grow* is a natural-kind *verb*, and precisely the sort of account we have been presenting applies to *it*.

Another misunderstanding that should be avoided is the following: to take the account we have developed as implying that the members of the extension of a natural-kind word necessarily *have* a common hidden

structure. It could have turned out that the bits of liquid we call 'water' had *no* important common physical characteristics *except* the superficial ones. In that case the necessary and sufficient condition for being 'water' would have been possession of sufficiently many of the superficial characteristics.

Incidentally, the last statement does not imply that water could have failed to have a hidden structure (or that water could have been anything but H_2O). When we say that it could have *turned out* that water had no hidden structure what we mean is that a liquid with no hidden structure (i.e. many bits of different liquids, with nothing in common *except* superficial characteristics) could have looked like water, tasted like water, and have filled the lakes, etc., that are actually full of water. In short, we could have been in the same epistemological situation with respect to a liquid with no hidden structure as we were actually with respect to water at one time. Compare Kripke on the 'lectern made of ice' (Kripke, 1972).

There are, in fact, almost continuously many cases. Some diseases, for example, have turned out to have no hidden structure (the only thing the paradigm cases have in common is a cluster of symptoms), while others have turned out to have a common hidden structure in the sense of an etiology (e.g. tuberculosis). Sometimes we still don't know; there is a controversy still raging about the case of multiple sclerosis.

An interesting case is the case of *jade*. Although the Chinese do not recognize a difference, the term 'jade' applies to two minerals: jadeite and nephrite. Chemically, there is a marked difference. Jadeite is a combination of sodium and aluminum. Nephrite is made of calcium, magnesium, and iron. These two quite different microstructures produce the same unique textural qualities!

Coming back to the Twin Earth example, for a moment; if H_2O and XYZ had both been plentiful on Earth, then we would have had a case similar to the jadeite/nephrite case: it would have been correct to say that there were *two kinds of 'water'*. And instead of saying that 'the stuff on Twin Earth turned out not to really be water', we would have to say 'it turned out to be the *XYZ kind of water*'.

To sum up: if there is a hidden structure, then generally it determines what it is to be a member of the natural kind, not only in the actual world, but in all possible worlds. Put another way, it determines what we can and cannot counterfactually suppose about the natural kind ('water could have all been vapor?' yes/'water could have been XYZ' no). But the local water, or whatever, may have two or more hidden structures – or so many that 'hidden structure' becomes irrelevant, and superficial characteristics become the decisive ones.

Other words

So far we have only used natural-kind words as examples; but the points we have made apply to many other kinds of words as well. They apply to the great majority of all nouns, and to other parts of speech as well.

Let us consider for a moment the names of artifacts – words like 'pencil', 'chair', 'bottle', etc. The traditional view is that these words are certainly defined by conjunctions, or possibly clusters, of properties. Anything with all of the properties in the conjunction (or sufficiently many of the properties in the cluster, on the cluster model) is necessarily a *pencil*, *chair*, *bottle*, or whatever. In addition, some of the properties in the cluster (on the cluster model) are usually held to be *necessary* (on the conjunction-of-properties model, *all* of the properties in the conjunction are necessary). *Being an artifact* is supposedly necessary, and belonging to a kind with a certain standard purpose – e.g. 'pencils are artifacts', and 'pencils are standardly intended to be written with' are supposed to be necessary. Finally, this sort of necessity is held to be *epistemic* necessity – in fact, analyticity.

Let us once again engage in science fiction. This time we use an example devised by Rogers Albritton. Imagine that we someday discover that *pencils are organisms*. We cut them open and examine them under the electron microscope, and we see the almost invisible tracery of nerves and other organs. We spy upon them, and we see them spawn, and we see the offspring grow into full-grown pencils. We discover that these organisms are not imitating other (artifactual) pencils – there are not and never were any pencils except these organisms. It is strange, to be sure, that there is *lettering* on many of these organisms – e.g. BONDED *Grants* DELUXE made in U.S.A. No. 2. – perhaps they are intelligent organisms, and this is their form of camouflage. (We also have to explain why no one ever attempted to manufacture pencils, etc., but this is clearly a possible world, in some sense.)

If this is conceivable, and I agree with Albritton that it is, then it is epistemically possible that *pencils could turn out to be organisms*. It follows that *pencils are artifacts* is not epistemically necessary in the strongest sense and, *a fortiori*, not analytic.

Let us be careful, however. Have we shown that there is a possible world in which pencils are organisms? I think not. What we have shown is that there is a possible world in which certain organisms are the *epistemic counterparts* of pencils (the phrase is Kripke's). To return to the device of Twin Earth: imagine this time that pencils on Earth are just what we think they are, artifacts manufactured to be written with, while 'pencils' on Twin Earth are organisms à la Albritton. Imagine,

further, that this is totally unsuspected by the Twin Earthians – they have exactly the beliefs about 'pencils' that we have about pencils. When we discovered this, we would not say: 'some pencils are organisms'. We would be far more likely to say: 'the things on Twin Earth that pass for pencils aren't really pencils. They're really a species of organism'.

Suppose now the situation to be as in Albritton's example both on Earth and on Twin Earth. Then we would say 'pencils are organisms'. Thus, whether the 'pencil-organisms' on Twin Earth (or in another possible universe) are really *pencils* or not is a function of whether or not the *local* pencils are organisms or not. If the local pencils are just what we think they are, then a possible world in which there are pencil-organisms is *not* a possible world in which *pencils are organisms*; there are *no* possible worlds in which pencils are organisms in this case (which is, of course, the actual one). That pencils are artifacts *is* necessary in the sense of true in all possible worlds – metaphysically necessary. But it doesn't follow that it's epistemically necessary.

It follows that 'pencil' is not *synonymous* with any description – not even loosely synonymous with a *loose* description. When we use the word 'pencil', we intend to refer to whatever has the same *nature* as the normal examples of the local pencils in the actual world. 'Pencil' is just as *indexical* as 'water' or 'gold'.

In a way, the case of pencils turning out to be organisms is complementary to the case we discussed some years ago (see my chapter 15, volume 1) of cats turning out to be robots (remotely controlled from Mars). In Katz (forthcoming), Katz argues that we misdescribed this case: that the case should rather be described as its *turning out that there are no cats in this world*. Katz admits that we might *say* 'Cats have turned out not to be animals, but robots'; but he argues that this is a semantically deviant sentence which is glossed as 'the things I am referring to as "cats" have turned out not to be animals, but robots'. Katz's theory is bad linguistics, however. First of all, the explanation of how it is we can *say* 'Cats are robots' is simply an all-purpose explanation of how we can say *anything*. More important, Katz's theory predicts that 'Cats are robots' is *deviant*, while 'There are no cats in the world' is nondeviant, in fact standard, in the case described. Now then, I don't deny that there *is* a case in which 'There are not (and never were) any cats in the world' would be standard: we might (speaking epistemically) discover that we have been suffering from a collective hallucination. ('Cats' are like pink elephants.) But in the case I described, 'Cats have turned out to be robots remotely controlled from Mars' is surely nondeviant, and 'There are no cats in the world' is highly deviant.

Incidentally, Katz's account is not only bad linguistics; it is also bad

as a rational reconstruction. The reason we *don't* use 'cat' as synonymous with a description is surely that we know enough about cats to know that they do have a hidden structure, and it is good scientific methodology to use the name to refer rigidly to the things that possess that hidden structure, and not to whatever happens to satisfy some description. Of course, if we *knew* the hidden structure we could frame a description in terms of *it*; but we don't at this point. In this sense the use of natural-kind words reflects an important fact about our relation to the world: we know that there are kinds of things with common hidden structure, but we don't yet have the knowledge to describe all those hidden structures.

Katz's view has more plausibility in the 'pencil' case than in the 'cat' case, however. We think we *know* a necessary and sufficient condition for being a *pencil*, albeit a vague one. So it is possible to make 'pencil' synonymous with a loose description. We *might* say, in the case that 'pencils turned out to be organisms' *either* 'Pencils have turned out to be organisms' *or* 'There are no pencils in the world' – i.e. we might use 'pencil' either as a natural-kind word or as a 'one-criterion' word.†

On the other hand, we might doubt that there *are* any true one-criterion words in natural language, apart from stipulative contexts. Couldn't it turn out that pediatricians aren't doctors but Martian spies? Answer 'yes', and you have abandoned the synonymy of 'pediatrician' and 'doctor specializing in the care of children'. It seems that there is a strong tendency for words which are introduced as 'one-criterion' words to develop a 'natural kind' sense, with all the concomitant rigidity and indexicality. In the case of artifact-names, this natural-kind sense seems to be the predominant one.

(There is a joke about a patient who is on the verge of being discharged from an insane asylum. The doctors have been questioning him for some time, and he has been giving perfectly sane responses. They decide to let him leave, and at the end of the interview one of the doctors inquires casually, 'What do you want to be when you get out?' 'A teakettle'. The joke would not be intelligible if it were literally inconceivable that a person could be a teakettle.)

There are, however, words which retain an almost pure one-criterion character. These are words whose meaning derives from a transformation: *hunter = one who hunts*.

Not only does the account given here apply to most nouns, but it also applies to other parts of speech. Verbs like 'grow', adjectives like 'red', etc., all have indexical features. On the other hand, some syncategorematic words seem to have more of a one-criterion character. 'Whole', for example, can be explained thus: *The army surrounded the town* could be true even if the *A* division did not take part. *The whole army surrounded the town* means every part of the army (of the relevant kind, e.g. the *A* Division) took part in the action signified by the verb.†

† The idea of a 'one-criterion' word, and a theory of analyticity based on this notion, appears in chapter 2 in this volume.

Meaning

Let us now see where we are with respect to the notion of meaning. We have now seen that the extension of a term is not fixed by a concept that the individual speaker has in his head, and this is true both because extension is, in general, determined *socially* – there is division of linguistic labor as much as of 'real' labor – and because extension is, in part, determined *indexically*. The extension of our terms depends upon the actual nature of the particular things that serve as paradigms,‡ and this actual nature is not, in general, fully known to the speaker. Traditional semantic theory leaves out only two contributions to the determination of extension – the contribution of society and the contribution of the real world!

We saw at the outset that meaning cannot be identified with extension. Yet it cannot be identified with 'intension' either, if intension is something like an individual speaker's *concept*. What are we to do?

There are two plausible routes that we might take. One route would be to retain the identification of meaning with concept and pay the price of giving up the idea that meaning determines extension. If we followed this route, we might say that 'water' has the same *meaning* on Earth and on Twin Earth, but a different *extension*. (Not just a different *local* extension but a different *global* extension. The *XYZ* on Twin Earth isn't in the extension of the tokens of 'water' that I utter, but it is in the extension of the tokens of 'water' that my *Doppelgänger* utters, and this isn't just because Twin Earth is far away from me, since molecules of H_2O are in the extension of the tokens of 'water' that I utter no matter how far away from me they are in space and time. Also, what I can counterfactually suppose water to be is different from what my *Dopelgänger* can counterfactually suppose 'water' to be.) While this is the correct route to take for an *absolutely* indexical word like 'I', it seems incorrect for the words we have been discussing. Consider 'elm' and 'beech', for example. If these are 'switched' on Twin Earth, then surely we would *not* say that 'elm' has the same meaning on Earth and Twin

† This example comes from an analysis by Anthony Kroch (in his M.I.T. doctoral dissertation, 1974, Department of Linguistics).

‡ I *don't* have in mind the Flewish notion of 'paradigm' in which any paradigm of a *K* is *necessarily* a *K* (in reality).

Earth, even if my *Doppelgänger's* stereotype of a beech (or an 'elm', as he calls it) is identical with my stereotype of an elm. Rather, we would say that 'elm' in my *Doppelgänger's* idiolect means *beech*. For this reason, it seems preferable to take a different route and identify 'meaning' with an ordered pair (or possibly an ordered *n-tuple*) of entities, *one of which is the extension*. (The other components of the, so to speak, 'meaning vector' will be specified later). Doing this makes it trivially true that *meaning determines extension* (i.e. difference in extension is *ipso facto* difference in meaning), but totally abandons the idea that if there is a difference in the meaning my *Doppelgänger* and I assign to a word, then there *must* be some difference in our concepts (or in our psychological state). Following this route, we can say that my *Doppelgänger* and I *mean something different* when we say 'elm', but this will not be an assertion about our psychological states. All this means is that the tokens of the word he utters have a different extension than the tokens of the word I utter; but this difference in extension is not a reflection of any difference in our individual linguistic competence considered in isolation.

If this is correct, and I think it is, then the traditional problem of meaning splits into two problems. The first problem is to account for the *determination of extension*. Since, in many cases, extension is determined socially and not individually, owing to the division of linguistic labor, I believe that this problem is properly a problem for sociolinguistics. Solving it would involve spelling out in detail exactly how the division of linguistic labor works. The so-called 'causal theory of reference', introduced by Kripke for proper names and extended by us to natural-kind words and physical-magnitude terms (in the preceding chapter), falls into this province. For the fact that, in many contexts, we assign to the tokens of a name that I utter whatever referent we assign to the tokens of the same name uttered by the person from whom I acquired the name (so that the reference is transmitted from speaker to speaker, starting from the speakers who were present at the 'naming ceremony', even though no fixed *description* is transmitted) is simply a special case of social cooperation in the determination of reference.

The other problem is to describe *individual competence*. Extension may be determined socially, in many cases, but we don't assign the standard extension to the tokens of a word *W* uttered by Jones *no matter how* Jones uses *W*. Jones has to have some particular ideas and skills in connection with *W* in order to play his part in the linguistic division of labor. Once we give up the idea that individual competence has to be so strong as to actually determine extension, we can begin to study it in a fresh frame of mind.

In this connection it is instructive to observe that nouns like 'tiger' or 'water' are very different from proper names. One can use the proper name 'Sanders' correctly without knowing anything about the referent except that he is called 'Sanders' – and even that may not be correct. ('Once upon a time, a very long time ago now, about last Friday, Winnie-the-Pooh lived in a forest all by himself under the name of Sanders.') But one cannot use the word tiger correctly, save *per accidens*, without knowing a good deal about tigers, or at least about a certain conception of tigers. In this sense concepts *do* have a lot to do with meaning.

Just as the study of the first problem is properly a topic in sociolinguistics, so the study of the second problem is properly a topic in psycholinguistics. To this topic we now turn.

Stereotypes and communication

Suppose a speaker knows that 'tiger' has a set of physical objects as its extension, but no more. If he possesses normal linguistic competence in other respects, then he could use 'tiger' in *some* sentences: for example, 'tigers have mass', 'tigers take up space', 'give me a tiger', 'is that a tiger?', etc. Moreover, the *socially determined* extension of 'tiger' in these sentences would be the standard one, i.e. the set of tigers. Yet we would not count such a speaker as 'knowing the meaning' of the word *tiger*. Why not?

Before attempting to answer this question, let us reformulate it a bit. We shall speak of someone as having *acquired* the word 'tiger' if he is able to use it in such a way that (1) his use passes muster (i.e. people don't say of him such things as 'he doesn't know what a tiger *is*', 'he doesn't know the meaning of the word "tiger"', etc.); and (2) his total way of being situated in the world and in his linguistic community is such that the socially determined extension of the word 'tiger' in his idiolect is the set of tigers. Clause (1) means, roughly, that speakers like the one hypothesized in the preceding paragraph don't count as having acquired the word 'tiger' (or whichever). We might speak of them, in some cases, as having *partially acquired* the word; but let us defer this for the moment. Clause (2) means that speakers on Twin Earth who have the same linguistic habits as we do, count as having acquired the word 'tiger' only if the extension of 'tiger' in their idiolect is the set of tigers. The burden of the preceding sections of this paper is that it does *not* follow that the extension of 'tiger' in Twin Earth dialect (or idiolects) is the set of tigers merely because their linguistic habits are the same as ours: the nature of Twin Earth 'tigers' is also relevant. (If Twin Earth organisms have a silicon chemistry, for example, then their 'tigers'

aren't really tigers, even if they look like tigers, although the linguistic habits of the lay Twin Earth speaker exactly correspond to those of Earth speakers.) Thus clause (2) means that in this case we have decided to say that Twin Earth speakers have not acquired our word 'tiger' (although they have acquired another word with the same spelling and pronunciation).

Our reason for introducing this way of speaking is that the question 'does he know the meaning of the word "tiger"?' is biassed in favor of the theory that acquiring a word is coming to possess a thing called its 'meaning'. Identify this thing with a concept, and we are back at the theory that a sufficient condition for acquiring a word is associating it with the right concept (or, more generally, being in the right psychological state with respect to it) – the very theory we have spent all this time refuting. So, henceforth, we will 'acquire' words, rather than 'learn their meaning'.

We can now reformulate the question with which this section began. The use of the speaker we described does not pass muster, although it is not such as to cause us to assign a nonstandard extension to the word 'tiger' in his idiolect. Why doesn't it pass muster?

Suppose our hypothetical speaker points to a snowball and asks, 'is that a tiger?'. Clearly there isn't much point in talking tigers with *him*. Significant communication requires that people know something of what they are talking about. To be sure, we hear people 'communicating' every day who clearly know nothing of what they are talking about; but the sense in which the man who points to a snowball and asks 'is that a tiger?' doesn't know anything about tigers is so far beyond the sense in which the man who thinks that Vancouver is going to win the Stanley Cup, or that the Vietnam War was fought to help the South Vietnamese, doesn't know what he is talking about as to boggle the mind. The problem of people who think that Vancouver is going to win the Stanley Cup, or that the Vietnam war was fought to help the South Vietnamese, is one that obviously cannot be remedied by the adoption of linguistic conventions; but not knowing what one is talking about in the second, mind-boggling sense can be and is prevented, near enough, by our conventions of language. What I contend is that speakers are *required* to know something about (stereotypical) tigers in order to count as having acquired the word 'tiger'; something about elm trees (or anyway, about the stereotype thereof) to count as having acquired the word 'elm'; etc.

This idea should not seem too surprising. After all, we do not permit people to drive on the highways without first passing some tests to determine that they have a *minimum* level of competence; and we do not dine with people who have not learned to use a knife and fork. The

linguistic community too has its minimum standards, with respect both to syntax and to 'semantics'.

The nature of the required minimum level of competence depends heavily upon both the culture and the topic, however. In our culture speakers are required to know what tigers look like (if they acquire the word 'tiger', and this is virtually obligatory); they are not required to know the fine details (such as leaf shape) of what an elm tree looks like. English speakers are ***required by their linguistic community*** to be able to tell tigers from leopards; they are not required to be able to tell elm trees from beech trees.

This could easily have been different. Imagine an Indian tribe, call it the Cheroquoi, who have words, say *uhaba'* and *wa'arabi* for elm trees and beech trees respectively, and who make it *obligatory* to know the difference. A Cheroquoi who could not recognize an elm would be said not to know what an *uhaba'* is, not to know the meaning of the word *uhaba'* (perhaps, not to know the word, or not to *have* the word); just as an English speaker who had no idea that tigers are striped would be said not to know what a tiger is, not to know the meaning of the word 'tiger' (of course, if he at least knows that tigers are large felines we might say he knows part of the meaning, or partially knows the meaning), etc. Then the translation of *uhaba'* as 'elm' and *wa'arabi* as 'beech' would, in our view, be only *approximately* correct. In this sense there is a real difficulty with radical translation,† but this is not the abstract difficulty that Quine is talking about.‡

What stereotypes are

I introduced the notion of a 'stereotype' in my lectures at the University of Washington and at the Minnesota Center for the Philosophy of Science in 1968. The subsequently published 'Is semantics possible?' (chapter 8 in this volume) follows up the argumentation, and in the present essay I want to introduce the notion again and to answer some questions that have been asked about it.

In ordinary parlance a 'stereotype' is a conventional (frequently malicious) idea (which may be wildly inaccurate) of what an X looks like or acts like or is. Obviously, I am trading on some features of the ordinary parlance. I am not concerned with malicious stereotypes (save where the language itself is malicious); but I am concerned with conventional ideas, which may be inaccurate. I am suggesting that just such

† The term is due to Quine (in *Word and Object*): it signifies translation without clues either from shared culture or cognates.

‡ For a discussion of the supposed impossibility of uniquely correct radical translation see chapter 9 in this volume.

a conventional idea is associated with 'tiger', with 'gold', etc., and, moreover, that this is the sole element of truth in the 'concept' theory.

In this view someone who knows what 'tiger' means (or, as we have decided to say instead, has acquired the word 'tiger') is ***required*** to know that ***stereotypical*** tigers are striped. More precisely, there is ***one*** stereotype of tigers (he may have others) which is required by the linguistic community as such; he is required to have this stereotype, and to know (implicitly) that it is obligatory. This stereotype must include the feature of stripes if his acquisition is to count as successful.

The fact that a feature (e.g. stripes) is included in the stereotype associated with a word *X* does not mean that it is an analytic truth that all *X*s have that feature, nor that most *X*s have that feature, nor that all normal *X*s have that feature, nor that some *X*s have that feature.† Three-legged tigers and albino tigers are not logically contradictory entities. Discovering that our stereotype has been based on nonnormal or unrepresentative members of a natural kind is not discovering a logical contradiction. If tigers lost their stripes they would not thereby cease to be tigers, nor would butterflies necessarily cease to be butterflies if they lost their wings.

(Strictly speaking, the situation is more complicated than this. It is possible to give a word like 'butterfly' a sense in which butterflies would cease to be butterflies if they lost their wings – through mutation, say. Thus one can find *a* sense of 'butterfly' in which it is analytic that 'butterflies have wings'. But the most important sense of the term, I believe, is the one in which the wingless butterflies would still be butterflies.)

At this point the reader may wonder what the value to the linguistic community of having stereotypes is, if the 'information' contained in the stereotype is not necessarily correct. But this is not really such a mystery. Most stereotypes do in fact capture features possessed by paradigmatic members of the class in question. Even where stereotypes go wrong, the way in which they go wrong sheds light on the contribution normally made by stereotypes to communication. The stereotype of gold, for example, contains the feature *yellow* even though chemically pure gold is nearly white. But the gold we see in jewelry is typically yellow (due to the presence of copper), so the presence of this feature in the stereotype is even useful in lay contexts. The stereotype associated with *witch* is more seriously wrong, at least if taken with existential import. Believing (with existential import) that witches enter into pacts with Satan, that they cause sickness and death, etc., facilitates communication only in the sense of facilitating communication internal to witch-

† This is argued in chapter 8.

theory. It does not facilitate communication in any situation in which what is needed is more agreement with the world than agreement with the theory of other speakers. (Strictly speaking, I am speaking of the stereotype as it existed in New England three hundred years ago; today that witches aren't *real* is itself part of the stereotype, and the baneful effects of witch-theory are thereby neutralized.) But the fact that our language has *some* stereotypes which impede rather than facilitate our dealings with the world and each other only points to the fact that we aren't infallible beings, and how could we be? The fact is that we could hardly communicate successfully if most of our stereotypes weren't pretty accurate as far as they go.

The 'operational meaning' of stereotypes

A trickier question is this: how far is the notion of stereotype 'operationally definable'. Here it is necessary to be extremely careful. Attempts in the physical sciences to *literally* specify operational definitions for terms have notoriously failed; and there is no reason the attempt should succeed in linguistics when it failed in physics. Sometimes Quine's arguments against the possibility of a theory of meaning seem to reduce to the demand for operational definitions in linguistics; when this is the case the arguments should be ignored. But it frequently happens that terms do have operational definitions not in the actual world but in idealized circumstances. Giving these 'operational definitions' has heuristic value, as idealization frequently does. It is only when we mistake operational definition for more than convenient idealization that it becomes harmful. Thus we may ask: what is the 'operational meaning' of the statement that a word has such and such a stereotype, without supposing that the answer to this question counts as a theoretical account of what it is to be a stereotype.

The theoretical account of what it is to be a stereotype proceeds in terms of the notion of *linguistic obligation*; a notion which we believe to be fundamental to linguistics and which we shall not attempt to explicate here. What it means to say that being striped is part of the (linguistic) stereotype of 'tiger' is that it is *obligatory* to acquire the information that stereotypical tigers are striped if one acquires 'tiger', in the same sense of 'obligatory' in which it is obligatory to indicate whether one is speaking of lions in the singular or lions in the plural when one speaks of lions in English. To describe an idealized experimental test of this hypothesis is not difficult. Let us introduce a person whom we may call the linguist's *confederate*. The confederate will be (or pretend to be) an adult whose command of English is generally excellent, but who

for some reason (raised in an alien culture? brought up in a monastery?) has totally failed to acquire the word 'tiger'. The confederate will say the word 'tiger' or, better yet, point to it (as if he wasn't sure how to pronounce it), and ask 'what does this word mean?' or 'what is this?' or some such question. Ignoring all the things that go wrong with experiments in practice, what our hypothesis implies is that informants should typically tell the confederate that tigers are, *inter alia*, striped.

Instead of relying on confederates, one might expect the linguist to study children learning English. But children learning their native language aren't taught it nearly as much as philosophers suppose; they learn it but they aren't taught it, as Chomsky has emphasized. Still, children do sometimes ask such questions as 'what is a tiger?' and our hypothesis implies that in these cases too informants should tell them, *inter alia*, that tigers are striped. But one problem is that the informants are likely to be parents, and there are the vagaries of parental time, temper, and attention to be allowed for.

It would be easy to specify a large number of additional 'operational' implications of our hypothesis, but to do so would have no particular value. The fact is that we are fully competent speakers of English ourselves, with a devil of a good sense of what our linguistic obligations are. Pretending that we are in the position of Martians with respect to English is not the route to methodological clarity; it was, after all, only when the operational approach was abandoned that transformational linguistics blossomed into a handsome science.

Thus if anyone were to ask me for the meaning of 'tiger', I know perfectly well what I would tell him. I would tell him that tigers were feline, something about their size, that they are yellow with black stripes, that they (sometimes) live in the jungle, and are fierce. Other things I might tell him too, depending on the context and his reason for asking; but the above items, save possibly for the bit about the jungle, I would regard it as *obligatory* to convey. I don't have to experiment to know that this is what I regard it as obligatory to convey, and I am sure that approximately this is what other speakers regard it as obligatory to convey too. Of course, there is some variation from idiolect to idiolect; the feature of having stripes (apart from figure–ground relations, e.g. are they black stripes on a yellow ground, which is the way I see them, or yellow stripes on a black ground?) would be found in all normal idiolects, but some speakers might regard the information that tigers (stereotypically) inhabit jungles as obligatory, while others might not. Alternatively, some features of the stereotype (big–cat–hood, stripes) might be regarded as obligatory, and others as *optional*, on the model of certain syntactical features. But we shall not pursue this possibility here.

Quine's 'Two dogmas' revisited

In 'Two dogmas of empiricism' Quine launched a powerful and salutory attack on the currently fashionable analytic–synthetic distinction. The distinction had grown to be a veritable philosophical man-eater: analytic *equalling* necessary *equalling* unrevisable in principle *equalling* whatever truth the individual philosopher wished to explain away. But Quine's attack itself went too far in certain respects; some limited class of analytic sentences can be saved, we feel (see chapter 2). More importantly, the attack was later construed, both by Quine himself and by others, as implicating the whole notion of meaning in the downfall of the analytic–synthetic distinction. While we have made it clear that we agree that the traditional notion of meaning has serious troubles, our project in this paper is constructive, not destructive. We come to revise the notion of meaning, not to bury it. So it will be useful to see how Quine's arguments fare against our revision.

Quine's arguments against the notion of analyticity can basically be reduced to the following: that no behavioral significance can be attached to the notion. His argument (again simplifying somewhat) was that there were, basically, only two candidates for a behavioral index of analyticity, and both are totally unsatisfactory, although for different reasons. The first behavioral index is *centrality*: many contemporary philosophers call a sentence analytic if, in effect some community (say, Oxford dons) holds it immune from revision. But, Quine persuasively argues, maximum immunity from revision is no exclusive prerogative of analytic sentences. Sentences expressing fundamental laws of physics (e.g. the conservation of energy) may well enjoy maximum behavioral immunity from revision, although it would hardly be customary or plausible to classify them as analytic. Quine does not, however, rely on the mere implausibility of classifying all statements that we are highly reluctant to give up as analytic; he points out that 'immunity from revision' is, in the actual history of science, a *matter of degree*. There is no such thing, in the actual practice of rational science, as *absolute* immunity from revision. Thus to identify analyticity with immunity from revision would alter the notion in two fundamental ways: analyticity would become a matter of degree, and there would be no such thing as an absolutely analytic sentence. This would be such a departure from the classical Carnap-Ayer-*et al.* notion of analyticity that Quine feels that if *this* is what we mean to talk about, then it would be less misleading to introduce a different term altogether, say, *centrality*.

The second behavioral index is *being called 'analytic'*. In effect, some philosophers take the hallmark of analyticity to be that trained inform-

ants (say, Oxford dons) *call* the sentence analytic. Variants of this index are: that the sentence be deducible from the sentences in a finite list at the top of which someone who bears the ancestral of the graduate-student relation to Carnap has printed the words 'Meaning Postulate'; that the sentence be obtainable from a theorem of logic by substituting synonyms for synonyms. The last of these variants looks promising, but Quine launches against it the question, 'what is the criterion of synonymy?'. One possible criterion might be that words W_1 and W_2 are synonymous if and only if the biconditional (x) (x is in the extension of $W_1 \equiv x$ is in the extension of W_2) is *analytic*; but this leads us right back in a circle. Another might be that words W_1 and W_2 are synonymous if and only if trained informants *call* them synonymous; but this is just our second index in a slightly revised form. A promising line is that words W_1 and W_2 are synonymous if and only if W_1 and W_2 are interchangeable (i.e. the words can be switched) *salva veritate* in all contexts of a suitable class. But Quine convincingly shows that this proposal too leads us around in a circle. Thus the second index reduces to this: a sentence is analytic if either it or some expression, or sequence of ordered pairs of expressions, or set of expressions, related to the sentence in certain specified ways, lies in a class to all the members of which trained informants apply a certain *noise*: either the *noise* ANALYTIC, or the *noise* MEANING POSTULATE, or the *noise* SYNONYMOUS. Ultimately, this proposal leaves 'analytic', etc., *unexplicated noises*.

Although Quine does not discuss this explicitly, it is clear that taking the intersection of the two unsatisfactory behavioral indexes would be no more satisfactory; explicating the analyticity of a sentence as consisting in centrality *plus* being called ANALYTIC is just saying that the analytic sentences are a subclass of the central sentences without in any way telling us wherein the exceptionality of the subclass consists. In effect, Quine's conclusion is that analyticity is either centrality misconceived or it is nothing.

In spite of Quine's forceful argument, many philosophers have gone on abusing the notion of analyticity, often confusing it with a supposed highest degree of centrality. Confronted with Quine's alternatives, they have elected to identify analyticity with centrality, and to pay the price – the price of classifying such obviously synthetic-looking sentences as 'space has three dimensions' as analytic, and the price of undertaking to maintain the view that there is, after all, such a thing as absolute unrevisability in science in spite of the impressive evidence to the contrary. But this line can be blasted by coupling Quine's argument with an important argument of Reichenbach's.

Reichenbach (Reichenbach, 1965, p. 31) showed that there exists a *set*

of principles each of which Kant would have regarded as synthetic *a priori*, but whose conjunction is incompatible with the principles of special relativity and general covariance. (These include normal induction, the continuity of space, and the Euclidean character of space.) A Kantian can consistently hold on to Euclidean geometry come what may; but then experience may force him to give up normal induction or the continuity of space. Or he may hold on to normal induction and the continuity of space come what may; but then experience may force him to give up Euclidean geometry (this happens in the case that physical space is not even homeomorphic to any Euclidean space). In his article in Schilpp (1951) Reichenbach gives essentially the same argument in a slightly different form.

Applied to our present context, what this shows is that there are principles such that philosophers fond of the overblown notion of analyticity, and in particular philosophers who identify analyticity with (maximum) unrevisability, would classify then as analytic, but whose conjunction has testable empirical consequences. Thus either the identification of analyticity with centrality must be given up once and for all, or one must give up the idea that analyticity is closed under conjunction, or one must swallow the unhappy consequence that an analytic sentence can have testable empirical consequences (and hence that an *analytic* sentence might turn out to be *empirically false*).

It is no accident, by the way, that the sentences that Kant would have classified as synthetic *a priori* would be classified by these latter-day empiricists as analytic; their purpose in bloating the notion of analyticity was precisely to dissolve Kant's problem by identifying *apriority* with analyticity and then identifying analyticity in turn with truth by convention. (This last step has also been devastingly criticized by Quine, but discussion of it would take us away from our topic.)

Other philosophers have tried to answer Quine by distinguishing between *sentences* and *statements*: all *sentences* are revisable, they agree, but some *statements* are not. Revising a sentence is not changing our mind about the statement formerly expressed by that sentence just in case the sentence (meaning the syntactical object together with its meaning) after the revision is, in fact, not synonymous with the sentence prior to the revision, i.e. just in case the revision is a case of meaning change and not change of theory. But (1) this reduces at once to the proposal to explicate analyticity in terms of synonymy; and (2) if there is one thing that Quine has decisively contributed to philosophy, it is the realization that meaning change and theory change cannot be sharply separated. We do not agree with Quine that meaning change cannot be defined at all, but it does not follow that the dichotomy

'meaning change *or* theory change' is tenable. Discovering that we live in a non-Euclidean world *might* change the meaning of 'straight line' (this would happen in the – somewhat unlikely – event that something like the parallels postulate was part of the stereotype of straightness); but it would not be a *mere* change of meaning. In particular it would not be a change of *extension*: thus it would not be right to say that the parallels postulate was 'true in the former sense of the words'. From the fact that giving up a sentence S would involve meaning change, it does not follow that S is *true*. Meanings may not fit the world; and meaning change can be forced by empirical discoveries.

Although we are not, in this paper, trying to explicate a notion of analyticity, we are trying to explicate a notion that might seem closely related, the notion of meaning. Thus it might seem that Quine's arguments would also go against our attempt. Let us check this out.

In our view there is a perfectly good sense in which being striped is part of the meaning of 'tiger'. But it does not follow, in our view, that 'tigers are striped' is analytic. If a mutation occurred, all tigers might be albinos. Communication presupposes that I have a stereotype of tigers which includes stripes, and that you have a stereotype of tigers which includes stripes, and that I know that your stereotype includes stripes, and that you know that my stereotype includes stripes, and that you know that I know...(and so on, à la Grice, forever). But it does not presuppose that any particular stereotype be *correct*, or that the majority of our stereotypes remain correct forever. Linguistic obligatoriness is not supposed to be an index of unrevisability or even of truth; thus we can hold that 'tigers are striped' is part of the meaning of 'tiger' without being trapped in the problems of analyticity.

Thus Quine's arguments against identifying analyticity with centrality are not arguments against identifying a feature's being 'part of the meaning' of X with its being obligatorily included in the stereotype of X. What of Quine's 'noise' argument?

Of course, evidence concerning what people *say*, including explicit metalinguistic remarks, is important in 'semantics' as it is in syntax. Thus, if a speaker points to a *clam* and asks 'is that a tiger?' people are likely to guffaw. (When they stop laughing) they might say 'he doesn't know the meaning of "tiger"', or 'he doesn't know what tigers are'. Such comments can be helpful to the linguist. But we are not *defining* the stereotype in terms of such comments. To say that being 'big-cat-like' is part of the meaning of tiger is not merely to say that application of 'tiger' to something which is not big-cat-like (and also not a tiger) would provoke certain *noises*. It is to say that speakers acquire the information that 'tigers are (stereotypically) big-cat-like' as they

acquire the word 'tiger' and that they feel an obligation to guarantee that those to whom they teach the use of the word do likewise. Information about the minimum skills required for entry into the linguistic community is significant information; no circularity of the kind Quine criticized appears here.

Radical translation

What our theory does not do, by itself at any rate, is solve Quine's problem of 'radical translation' (i.e. translation from an alien language/culture). We cannot translate our hypothetical Cheroquoi into English by matching stereotypes, just because finding out what the stereotype of, say, *wa'arabi* is involves translating Cheroquoi utterances. On the other hand, the constraint that each word in Cheroquoi should match its image in English under the translation-function as far as stereotype is concerned (or approximately match, since in many cases exact matching may not be attainable), places a severe *constraint* on the translation-function. Once we have succeeded in translating the basic vocabulary of Cheroquoi, we can start to elicit stereotypes, and these will serve both to constrain future translations and to check the internal correctness of the piece of the translation-function already constructed.

Even where we can determine stereotypes (relative, say, to a tentative translation of 'basic vocabulary'), these do not suffice, in general, to determine a unique translation. Thus the German words *Ulme* and *Buche* have the same stereotype as elm; but *Ulme* means 'elm' while *Buche* means 'beech'. In the case of German, the fact that *Ulme* and 'elm' are cognates could point to the correct translation (although this is far from foolproof – in general, cognate words are not synonymous); but in the case of Greek we have no such clue as to which of the two words ὀξύα, πτελέα means *elm* and which *beech*; we would just have to find a Greek who could tell elms from beeches (or *oxya* from *ptelea*). What this illustrates is that it may not be the *typical* speakers' dispositions to assent and dissent that the linguist must seek to discover; because of the division of linguistic labor, it is frequently necessary for the linguist to assess who are the experts with respect to *oxya*, or *wa'arabi*, or *gavagai*, or whatever, before he can make a guess at the socially determined extension of a word. Then this socially determined extension *and* the stereotype of the *typical* speaker, inexpert though he is, will *both* function as constraints upon the translation-function. Discovery that the stereotype of *oxya* is wildly different from the stereotype of 'elm' would disqualify the translation of *oxya* by 'elm' in all save the most extensional contexts; but the discovery that the *extension*

of *oxya* is not even approximately the class of elms would wipe out the translation altogether, in all contexts.

It will be noted that we have already enlarged the totality of facts counted as evidence for a translation-function beyond the ascetic base that Quine allows in *Word and Object*. For example, the fact that speakers say such-and-such when the linguist's 'confederate' points to the word *oxya* and asks 'what does this mean?' or 'what is this?' or whatever is not allowed by Quine (as something the linguist can 'know') on the ground that this sort of 'knowledge' presupposes already having translated the query 'what does this word mean?'. However, if Quine is willing to assume that one can *somehow* guess at the words which signify *assent* and *dissent* in the alien language, it does not seem at all unreasonable to suppose that one can somehow convey to a native speaker that one does not understand a word. It is not necessary that one discover a locution in the alien language which literally means 'what does this word mean?' (as opposed to: 'I don't understand this word', or 'this word is unfamiliar to me' or 'I am puzzled by this word', etc.). Perhaps just saying the word *oxya*, or whatever, with a tone of puzzlement would suffice. Why should *puzzlement* be less accessible to the linguist than *assent*?

Also, we are taking advantage of the fact that segmentation into *words* has turned out to be linguistically universal (and there even exist tests for word and morpheme segmentation which are independent of meaning). Clearly, there is no motivated reason for allowing the linguist to utter whole sentences and look for assent and dissent, while refusing to allow him to utter words and morphemes in a tone of puzzlement.

I repeat, the claim is not being advanced that enlarging the evidence base in this way solves the problem of radical translation. What it does is add further constraints on the class of admissible candidates for a correct translation. What I believe is that enlarging the class of constraints can determine a unique translation, or as unique a translation as we are able to get in practice. But constraints that go beyond linguistic theory proper will have to be used, in my opinion; there will also have to be constraints on what sorts of beliefs (and connections between beliefs, and connections of beliefs to the culture and the world) we can reasonably impute to people. Discussion of these matters will be deferred to another paper.

A critique of Davidsonian semantic theory

In a series of publications, Donald Davidson has put forward the interesting suggestion that a semantic theory of a natural language might be

modelled on what mathematical logicians call a ***truth definition*** for a formalized language. Stripped of technicalities, what this suggestion comes down to is that one might have a set of rules specifying (1) for each word, under what conditions that word is true of something (for words for which the concept of an extension makes sense; all other words are to be treated as syncategorematic); (2) for sentences longer than a single word, a rule is given specifying the conditions under which the sentence is true as a function of the way it is built up out of shorter sentences (counting words as if they were one-word sentences, e.g. 'snow' as 'that's snow'). The choice of one-word sentences as the starting point is my interpretation of what Davidson intends; in any case, he means one to start with a *finite* stock of *short* sentences for which truth conditions are to be laid down *directly*. The intention of (2) is not that there should be a rule for each sentence not handled under (1), since this would require an infinite number of rules, but that there should be a rule for each sentence *type*. For example, in a formalized language one of the rules of kind (2) might be: if S is $(S_1 \ \& \ S_2)$ for some sentences S_1, S_2, then S is true if and only if S_1, S_2, are both true.

It will be noticed that, in the example just given, the truth condition specified for sentences of the sentence type $(S_1 \ \& \ S_2)$ performs the job of specifying the meaning of '&'. More precisely, it specifies the meaning of the structure (—— & ——). This is the sense in which a truth definition can be a theory of meaning. Davidson's contention is that the *entire* theory of meaning for a natural language can be given in this form.

There is no doubt that rules of the type illustrated can give the meaning of some words and structures. The question is, what reason is there to think that the meaning of most words can be given in this way, let alone all?

The obvious difficulty is this: for many words, an extensionally correct truth definition can be given which is in no sense a theory of the meaning of the word. For example, consider '*Water' is true of* x *if and only if* x *is* H_2O. This is an extensionally correct truth definition for 'water' (strictly speaking, it is not a truth definition but a 'truth of' definition – i.e. a *satisfaction*-in-the-sense-of-Tarski definition, but we will not bother with such niceties here). At least it is extensionally correct if we ignore the problem that water with impurities is also called 'water', etc. Now, suppose most speakers don't *know* that water is H_2O. Then this formula in no way tells us anything about the *meaning* of 'water'. It might be of interest to a chemist, but it doesn't count as a theory of the meaning of the term 'water'. Or, it counts as a theory of the *extension* of the term 'water', but Davidson is promising us more than just that.

Davidson is quite well aware of this difficulty. His answer (in conversation, anyway) is that we need to develop a theory of *translation*. This he, like Quine, considers to be the real problem. Relativized to such a theory (relativitized to what we admittedly don't yet have), the theory comes down to this: we want a system of truth definitions which is simultaneously a system of translations (or approximate translations, if perfect translation is unobtainable). If we had a theory which specified what it is to be a good translation, then we could rule out the above truth definition for 'water' as uninteresting on the grounds that *x is* H_2O is not an acceptable translation or even near-translation of *x is water* (in a prescientific community), even if water = H_2O happens to be true.

This comes perilously close to saying that a theory of meaning is a truth definition plus a theory of meaning. (If we had ham and eggs we'd have ham and eggs – *if* we had ham and *if* we had eggs.) But this story suffers from worse than promissoriness, as we shall see.

A second contention of Davidson's is that the theory of translation that we don't yet have is necessarily a theory whose basic units are *sentences* and not *words* on the grounds that our *evidence* in linguistics necessarily consists of assent and dissent from sentences. Words can be handled, Davidson contends, by treating them as sentences ('water' as 'that's water', etc.).

How does this ambitious project of constructing a theory of meaning in the form of a truth definition constrained by a theory of translation tested by 'the only evidence we have', speakers' dispositions to use sentences, fare according to the view we are putting forward here?

Our answer is that the theory cannot succeed in principle. In special cases, such as the word 'and' in its truth-functional sense, a truth definition (strictly speaking, a clause in what logicians call a 'truth definition' – the sum total of all the clauses is the inductive definition of 'truth' for the particular language) can give the meaning of the word or structure because the stereotype associated with the word (if one wants to speak of a stereotype in the case of a word like 'and') is so strong as to actually constitute a necessary and sufficient condition. If all words were like 'and' and 'bachelor' the program could succeed. And Davidson certainly made an important contribution in pointing out that linguistics has to deal with inductively specified truth conditions. But in the great majority of words, the requirements of a theory of truth and the requirements of a theory of meaning are mutually incompatible, at least in the English–English case. But the English–English case – the case in which we try to provide a significant theory of the meaning of English words which is itself couched in English – is surely the basic one.

The problem is that in general the only expressions which are both coextensive with *X* and have roughly the same stereotype as *X* are expressions containing *X* itself. If we rule out such truth definitions (strictly speaking, clauses, but I shall continue using 'truth definition' both for individual clauses and for the whole system of clauses, for simplicity) as

'X *is water*' *is true if and only if* X *is water*

on the grounds that they don't say anything about the meaning of the word 'water', and we rule out such truth definitions as

'X *is water*' *is true if and only if* X *is* H_2O

on the grounds that what they say is wrong as a description of the *meaning* of the word 'water', then we shall be left with nothing.

The problem is that we want

W *is true of* x *if and only if* ——

to satisfy the conditions that (1) the clause be extensionally correct (where —— is to be thought of as a condition containing '*x*', e.g. '*x* is H_2O'); (2) that —— be a *translation* of *W* – on our theory, this would mean that the stereotype associated with *W* is approximately the same as the stereotype associated with ——; (3) that —— not contain *W* itself, or syntactic variants of *W*. If we take *W* to be, for example, the word 'elm', then there is absolutely no way to fulfill all three conditions simultaneously. Any condition of the above form that does not contain 'elm' and that is extensionally correct will contain a —— that is absolutely terrible as a *translation* of 'elm'.

Even where the language contains two exact synonyms, the situation is little better. Thus

'*Heather*' *is true of* x *if and only if* x *is gorse*

is true, and so is

'*Gorse*' *is true of* x *if and only if* x *is heather*

– *this* is a *theory* of the *meaning* of 'gorse' and 'heather'?

Notice that the condition (3) is precisely what logicians do *not* impose on *their* truth definitions.

'*Snow is white*' *is true if and only if snow is white*

is the paradigm of a truth definition in the logician's sense. But logicians are trying to give the extension of 'true' with respect to a particular language, not the meaning of 'snow is white'. Tarski would have gone so far as to claim he was giving the *meaning* (and not just the extension)

of 'true'; but he would never have claimed he was saying *anything* about the meaning of 'snow is white'.

It may be that what Davidson really thinks is that theory of meaning, in any serious sense of the term, is impossible, and that all that is possible is to construct translation-functions. If so, he might well think that the only 'theory of meaning' possible for English is one that says '"elm" is true of x if and only if x is an elm', '"water" is true of x if and only if x is water', etc., and only rarely something enlightening like 'S_1 & S_2 is true if and only if S_1, S_2 are both true'. But if Davidson's 'theory' is just Quinine skepticism under the disguise of a positive contribution to the study of meaning, then it is a bitter pill to swallow.

The contention that the only evidence available to the linguist is speakers' dispositions with respect to whole sentences is, furthermore, vacuous on one interpretation, and plainly false on the interpretation on which it is not vacuous. If dispositions to say certain things *when queried about individual words or morphemes or syntactic structures* are included in the notion of dispositions to use sentences, then the restriction to dispositions to use sentences seems to rule out nothing whatsoever. On the non-vacuous interpretation, what Davidson is saying is that the linguist cannot have access to such data as what informants (including the linguist himself) say when asked the meaning of a word or morpheme or syntactic structure. No reason has ever been given why the linguist cannot have access to such data, and it is plain that actual linguists place heavy reliance on informants' testimony about such matters, in the case of an alien language, and upon their own intuitions as native speakers, when they are studying their native languages. In particular, when we are trying to translate a whole sentence, there is no reason why we should not be guided by our knowledge of the syntactic and semantic properties of the constituents of that sentence, including the deep structure. As we have seen, there are procedures for gaining information about individual constituents. It is noteworthy that the procedure that Quine and Davidson claim is the only *possible* one – going from whole sentences to individual words – is the *opposite* of the procedure upon which every success ever attained in the study of natural language has been based.

Critique of California semantics

I wish now to consider an approach to semantic theory pioneered by the late Rudolf Carnap. Since I do not wish to be embroiled in textual questions, I will not attribute the particular form of the view I am going

to describe to any particular philosopher but will simply refer to it as 'California semantics'.

We assume the notion of a *possible world*. Let f be a function defined on the 'space' of all possible worlds whose value $f(x)$ at any possible world x is always a subset of the set of entities in x. Then f is called an *intension*. A term T has meaning for a speaker X if X associates T with an intension f_T. The term T is *true of* an entity e in a possible world x if and only if e belongs to the set $f(x)$. Instead of using the term 'associated', Carnap himself tended to speak of 'grasping' intensions; but, clearly, what was intended was not just that X 'grasp' the intension f, but that he grasp *that* f is the intension *of* T – i.e. that he *associate* f with T in some way.

Clearly this picture of what it is to understand a term disagrees with the story we tell in this paper. The reply of a California semanticist would be that California sementics is a description of an *ideal* language; that actual language is *vague*. In other words, a term T in actual language does not have a single precise intension; it has a set – possibly a fuzzy set – of intensions. Nevertheless, the first step in the direction of describing natural language is surely to study the idealization in which each term T has exactly one intension.

(In his book *Meaning and Necessity*, Carnap employs a superficially different formulation: an intension is simply a *property*. An entity e belongs to the extension of a term T just in case e has whichever property is the intension of T. The later formulation in terms of functions f as described above avoids taking the notion of *property* as primitive.)

The first difficulty with this position is the use of the totally unexplained notion of *grasping* an intension (or, in our reformulation of the position, *associating* an intension with a term). Identifying intensions with set-theoretic entities f provides a 'concrete' realization of the notion of intension in the current mathematical style (relative to the notions of possible world and set), but at the cost of making it very difficult to see how anyone could have an intension in his mind, or what it is to think about one or 'grasp' one or 'associate' one with anything. It will not do to say that thinking of an intension is using a word or functional substitute for a word (e.g. the analogue of a word in 'brain code', if, as seems likely, the brain 'computes' in a 'code' that has analogies to and possibly borrowings from language; or a thought form such as a picture or a private symbol, in cases where such are employed in thinking) which *refers* to the intension in question, since *reference* (i.e. being in the extension of a term) has just been defined in terms of *intension*. Although the characterization of what it is to think of an abstract entity such as a function or a property is certainly correct, in

the present context it is patently circular. But no noncircular characterization of this fundamental notion of the theory has ever been provided.

This difficulty is related to a general difficulty in the philosophy of mathematics pointed out by Paul Benacerraf (Benacerraf, 1973). Benacerraf has remarked that philosophies of mathematics tend to fall between two stools: either they account for what mathematical objects are and for the necessity of mathematical truth and fail to account for the fact that people can *learn* mathematics, can *refer to* mathematical objects, etc., or else they account for the latter facts and fail to account for the former. California semantics accounts for what intensions *are*, but provides no account that is not completely circular of how it is that we can 'grasp' them, associate them with terms, think about them, *refer to* them, etc.

Carnap may not have noticed this difficulty because of his Verificationism. In his early years Carnap thought of understanding a term as possessing the *ability to verify* whether or not any given entity falls in the extension of the term. In terms of intensions: 'grasping' an intension would amount, then, to possessing the ability to verify if an entity e in any possible world x belongs to $f(x)$ or not. Later Carnap modified this view, recognizing that, as Quine puts it, sentences face the tribunal of experience collectively and not individually. There is no such thing as the way of verifying that a term T is true of an entity, in general, independent of the context of a particular set of theories, auxiliary hypotheses, etc. Perhaps Carnap would have maintained that something like the earlier theory was correct for a limited class of terms, the so-called 'observation terms'. Our own view is that the verifiability theory of meaning is false both in its central idea and for observation terms, but we shall not try to discuss this here. At any rate, if one is *not* a verificationist, then it is hard to see California semantics as a theory at all, since the notion of *grasping* an intension has been left totally unexplained.

Second, if we assume that 'grasping an intension' (associating an intension with a term T) is supposed to be a *psychological state* (in the narrow sense), then California semantics is committed to both principles (1) and (2) that we criticized in the first part of this paper. It must hold that the psychological state of the speaker determines the intension of his terms which in turn determines the extension of his terms. It would follow that if two human beings are in the same total psychological state, then they necessarily assign the same extension to every term they employ. As we have seen, this is totally wrong for natural language. The reason this is wrong, as we saw above, is in part that extension is determined socially, not by individual competence alone. Thus California

semantics is committed to treating language as something private – to totally ignoring the linguistic division of labor. The extension of each term is viewed by this school as totally determined by something in the head of the individual speaker all by himself. A second reason this is wrong, as we also saw, is that most terms are *rigid*. In California semantics every term is treated as, in effect, a *description*. The *indexical* component in meaning – the fact that our terms refer to things which are similar, in certain ways, to things that we designate *rigidly*, to *these* things, to the stuff we call 'water', or whatever, *here* – is ignored.

But what of the defense that it is not actual language that the California semanticist is concerned with, but an idealization in which we 'ignore vagueness', and that terms in natural language may be thought of as associated with a set of intensions rather than with a single well-defined intension?

The answer is that an *indexical* word cannot be represented as a vague family of non-indexical words. The word 'I', to take the extreme case, is *indexical* but not *vague*. 'I' is not synonymous with a *description*; neither is it synonymous with a fuzzy set of descriptions. Similarly, if we are right, 'water' is synonymous neither with a description nor with a fuzzy set of descriptions (intensions).

Similarly, a word whose extension is fixed socially and not individually is not the same thing as a word whose extension is *vaguely* fixed individually. The reason my individual 'grasp' of 'elm tree' does not fix the extension of elm is not that the word is vague – if the problem were simple vagueness, then the fact that my concepts do not distinguish elms from beeches would imply that elms are beeches, as I use the term, or, anyway, borderline cases of beeches, and that beeches are elms, or borderline cases of elms. The reason is rather that the extension of 'elm tree' in my dialect is not fixed by what the average speaker 'grasps' or doesn't 'grasp' at all; it is fixed by the community, including the experts, through a complex cooperative process. A language which exemplifies the division of linguistic labor cannot be approximated successfully by a language which has vague terms and no linguistic division of labor. Cooperation isn't vagueness.

But, one might reply, couldn't one replace our actual language by a language in which (1) terms were replaced by coextensive terms which were *not* indexical (e.g. 'water' by 'H_2O', assuming 'H_2O' is not indexical); and (2) we eliminated the division of linguistic labor by making every speaker an expert on every topic?

We shall answer this question in the negative; but suppose, for a moment, the answer were 'yes'. What significance would this have? The 'ideal' language would in no sense be similar to our actual language;

nor would the difference be a matter of 'the vagueness of natural language'.

In fact, however, one can't carry out the replacement, for the very good reason that *all* natural-kind words and physical-magnitude words are indexical in the way we have described, 'hydrogen', and hence 'H_2O', just as much as 'water'. Perhaps 'sense data' terms are not indexical (apart from terms for the self), if such there be; but 'yellow' as a *thing* predicate is indexical for the same reason as 'tiger'; even if something *looks* yellow it may not *be* yellow. And it doesn't help to say that things that look yellow in normal circumstances (to normal perceivers) are yellow; 'normal' here has precisely the feature we called indexicality. There is simply no reason to believe that the project of reducing our language to nonindexical language could be carried out in principle.

The elimination of the division of linguistic labor might, I suppose, be carried out 'in principle'. But, if the division of linguistic labor is, as I conjectured, a linguistic universal, what interest is there in the possible existence of a language which lacks a constitutive feature of *human* language? A world in which every one is an expert on every topic is a world in which social laws are almost unimaginably different from what they now are. What is the *motivation* for taking such a world and such a language as the model for the analysis of *human* language?

Incidentally, philosophers who work in the tradition of California semantics have recently begun to modify the scheme to overcome just these defects. Thus it has been suggested that an intension might be a function whose arguments are not just possible worlds but, perhaps, a possible world, a speaker, and a nonlinguistic context of utterance. This would permit the representation of some kinds of indexicality and some kinds of division of linguistic labor in the model. As David Lewis develops these ideas, 'water', for example, would have the same *intension* (same function) on Earth and on Twin Earth, but a different extension. (In effect, Lewis retains assumption (1) from the discussion in the first part of this paper and gives up (2); we chose to give up (1) and retain (2).) There is no reason why the formal models developed by Carnap and his followers should not prove valuable when so modified. Our interest here has been not in the utility of the mathematical formalism but in the philosophy of language underlying the earlier versions of the view.

Semantic markers

If the approach suggested here is correct, then there is a great deal of scientific work to be done in (1) finding out what sorts of items can

appear in stereotypes; (2) working out a convenient system for representing stereotypes; etc. This work is not work that can be done by philosophical discussion, however. It is rather the province of linguistics and psycholinguistics. One idea that can, I believe, be of value is the idea of a *semantic marker*. The idea comes from the work of J. Katz and J. A. Fodor; we shall modify it somewhat here.

Consider the stereotype of 'tiger' for a moment. This includes such features as being an animal; being big-cat-like; having black stripes on a yellow ground (yellow stripes on a black ground?); etc. Now, there is something very special about the feature *animal*. In terms of Quine's notion of *centrality* or *unrevisability*, it is qualitatively different from the others listed. It is not impossible to imagine that tigers might not be animals (they might be robots). But spelling this out, they must always have been robots; we don't want to tell a story about the tigers being *replaced* by robots, because then the robots wouldn't be tigers. Or, if they weren't always robots, they must have *become* robots, which is even harder to imagine. If tigers are and always were robots, these robots mustn't be too 'intelligent', or else we may not have a case in which tigers aren't animals – we may, rather, have described a case in which some robots are animals. Best make them 'other directed' robots – say, have an operator on Mars controlling each motion remotely. Spelling this out, I repeat, is difficult, and it is curiously hard to think of the case to begin with, which is why it is easy to make the mistake of thinking that it is 'logically impossible' for a tiger *not* to be an animal. On the other hand, there is no difficulty in imagining an individual tiger that is not striped; it might be an albino. Nor is it difficult to imagine an individual tiger that doesn't look like a big cat: it might be horribly deformed. We can even imagine the whole species losing its stripes or becoming horribly deformed. But tigers ceasing to be animals? Great difficulty again!

Notice that we are not making the mistake that Quine rightly criticized, of attributing an absolute unrevisability to such statements as 'tigers are animals', 'tigers couldn't change from animals into something else and still be tigers'. Indeed, we can describe farfetched cases in which these statements would be given up. But we maintain that it is *qualitatively* harder to revise 'all tigers are animals' than 'all tigers have stripes' – indeed, the latter statement is not even true.

Not only do such features as 'animal', 'living thing', 'artifact', 'day of the week', 'period of time', attach with enormous centrality to the words 'tiger', 'clam', 'chair', 'Tuesday', 'hour'; but they also form part of a widely used and important *system of classification*. The centrality guarantees that items classified under these headings virtually

never have to be *re*classified; thus these headings are the natural ones to use as category-indicators in a host of contexts. It seems to me reasonable that, just as in syntax we use such markers as 'noun', 'adjective', and, more narrowly, 'concrete noun', 'verb taking a person as subject and an abstract object', etc., to classify words, so in semantics these category-indicators should be used as markers.

It is interesting that when Katz and Fodor originally introduced the idea of a semantic marker, they did not propose to exhaust the meaning – what we call the stereotype – by a list of such markers. Rather, the markers were restricted to just the category-indicators of high centrality, which is what we propose. The remaining features were simply listed as a 'distinguisher'. Their scheme is not easily comparable with ours, because they wanted the semantic markers *plus* the distinguisher to always give a necessary and sufficient condition for membership in the extension of the term. Since the whole thing – markers and distinguisher – were supposed to represent what every speaker implicitly knows, they were committed to the idea that every speaker implicitly knows of a necessary and sufficient condition for membership in the extension of 'gold', 'aluminum', 'elm' – which, as we have pointed out, is not the case. Later Katz went further and demanded that *all* the features constitute an *analytically* necessary and sufficient condition for membership in the extension. At this point he dropped the distinction between markers and distinguishers; if all the features have, so to speak, the infinite degree of centrality, why call some 'markers' and some 'distinguishers'? From our point of view, their original distinction between 'markers' and 'distinguisher' was sound – provided one drop the idea that the distinguisher provides (together with the markers) a necessary and sufficient condition, and the idea that any of this is a theory of *analyticity*. We suggest that the idea of a semantic marker is an important contribution, when taken as suggested here.

The meaning of 'meaning'

We may now summarize what has been said in the form of a proposal concerning how one might reconstruct the notion of 'meaning'. Our proposal is not the only one that might be advanced on the basis of these ideas, but it may serve to encapsulate some of the major points. In addition, I feel that it recovers as much of ordinary usage in common sense talk and in linguistics as one is likely to be able to conveniently preserve. Since, in my view something like the assumptions (I) and (II) listed in the first part of this paper are deeply embedded in ordinary meaning talk, and these assumptions are jointly inconsistent with the

facts, no reconstruction is going to be without some counter-intuitive consequences.

Briefly, my proposal is to define 'meaning' not by picking out an object which will be identified with the meaning (although that might be done in the usual set-theoretic style if one insists), but by specifying a normal form (or, rather, a *type* of normal form) for the description of meaning. If we know what a 'normal form description' of the meaning of a word should be, then, as far as I am concerned, we know what meaning *is* in any scientifically interesting sense.

My proposal is that the normal form description of the meaning of a word should be a finite sequence, or 'vector', whose components should certainly include the following (it might be desirable to have other types of components as well): (1) the syntactic markers that apply to the word, e.g. 'noun'; (2) the semantic markers that apply to the word, e.g. 'animal', 'period of time'; (3) a description of the additional features of the stereotype, if any; (4) a description of the extension.

The following convention is a part of this proposal: the components of the vector all represent a hypothesis about the individual speaker's competence, *except the extension.* Thus the normal form description for 'water' might be, in part:

SYNTACTIC MARKERS	SEMANTIC MARKERS	STEREOTYPE	EXTENSION
mass noun, concrete;	*natural kind; liquid;*	*colorless; transparent; tasteless; thirst-quenching; etc.*	H_2O (*give or take impurities*)

– this does *not* mean that knowledge of the fact that water is H_2O is being imputed to the individual speaker or even to the society. It means that (*we* say) the extension of the term 'water' as *they* (the speakers in question) use it is *in fact* H_2O. The objection 'who are *we* to say what the extension of *their* term is in fact' has been discussed above. Note that this is fundamentally an objection to the notion of *truth*, and that extension is a relative of truth and inherits the family problems.

Let us call two descriptions *equivalent* if they are the same except for the description of the extension, and the two descriptions are coextensive. Then, if the set variously described in the two descriptions is, *in fact*, the extension of the word in question, and the other components in the description are correct characterizations of the various aspects of competence they represent, *both* descriptions count as correct. Equivalent descriptions are both correct or both incorrect. This is another way of

making the point that, although we have to use a *description* of the extension to *give* the extension, we think of the component in question as being the *extension* (the *set*), not the description of the extension.

In particular the representation of the words 'water' in Earth dialect and 'water' in Twin Earth dialect would be the same except that in the last column the normal form description of the Twin Earth word 'water' would have XYZ and not H_2O. This means, in view of what has just been said, that we are ascribing the *same* linguistic competence to the typical Earthian/Twin Earthian speaker, but a different extension to the word, nonetheless.

This proposal means that we keep assumption (II) of our early discussion. Meaning determines extension – by construction, so to speak. But (I) is given up; the psychological state of the individual speaker does not determine 'what he means'.

In most contexts this will agree with the way we speak, I believe. But one paradox: suppose Oscar is a German–English bilingual. In our view, in his total collection of dialects, the words 'beech' and *Buche* are *exact synonyms*. The normal form descriptions of their meanings would be identical. But he might very well not know that they are synonyms! A speaker can have two synonyms in his vocabulary and not know that they are synonyms!

It is instructive to see how the failure of the apparently obvious 'if S_1 and S_2 are synonyms and Oscar understands both S_1 and S_2 then Oscar knows that S_1 and S_2 are synonyms' is related to the falsity of (I), in our analysis. Notice that if we had chosen to omit the extension as a component of the 'meaning-vector', which is David Lewis's proposal as I understand it, then we would have the paradox that 'elm' and 'beech' have the *same meaning* but different extensions!

On just about any materialist theory, believing a proposition is likely to involve processing some *representation* of that proposition, be it a sentence in a language, a piece of 'brain code', a thought form, or whatever. Materialists, and not only materialists, are reluctant to think that one can believe propositions *neat*. But even materialists tend to believe that, if one believes a proposition, *which* representation one employs is (pardon the pun) immaterial. If S_1 and S_2 are both representations that are *available* to me, then if I believe the proposition expressed by S_1 under the representation S_1, I must also believe it under the representation S_2 – at least, I must do this if I have any claim to rationality. But, as we have just seen, this isn't right. Oscar may well believe that *this* is a 'beech' (it has a sign on it that says 'beech'), but not believe or disbelieve that this is a '*Buche*'. It is not just that belief is a process involving representations; he believes the proposition (if one

wants to introduce 'propositions' at all) under one representation and not under another.

The amazing thing about the theory of meaning is how long the subject has been in the grip of philosophical misconceptions, and how strong these misconceptions are. Meaning has been identified with a necessary and sufficient condition by philosopher after philosopher. In the empiricist tradition, it has been identified with method of verification, again by philosopher after philosopher. Nor have these misconceptions had the virtue of exclusiveness; not a few philosophers have held that meaning = method of verification = necessary and sufficient condition.

On the other side, it is amazing how weak the grip of the facts has been. After all, what have been pointed out in this essay are little more than home truths about the way we use words and how much (or rather, how little) we actually know when we use them. My own reflection on these matters began after I published a paper in which I confidently maintained that the meaning of a word was 'a battery of semantical rules', (chapter 6 in this volume) and then began to wonder how the meaning of the common word 'gold' could be accounted for in this way. And it is not that philosophers had never considered such examples: Locke, for example, uses this word as an example and is not troubled by the idea that its meaning is a necessary and sufficient condition!

If there is a reason for both learned and lay opinion having gone so far astray with respect to a topic which deals, after all, with matters which are in everyone's experience, matters concerning which we all have more data than we know what to do with, matters concerning which we have, if we shed preconceptions, pretty clear intuitions, it must be connected to the fact that the grotesquely mistaken views of language which are and always have been current reflect two specific and very central philosophical tendencies: the tendency to treat cognition as a purely *individual* matter and the tendency to ignore the *world*, insofar as it consists of more than the individual's 'observations'. Ignoring the division of linguistic labor is ignoring the social dimension of cognition; ignoring what we have called the *indexicality* of most words is ignoring the contribution of the environment. Traditional philosophy of language, like much traditional philosophy, leaves out other people and the world; a better philosophy and a better science of language must encompass both.