THE JOURNAL OF PHILOSOPHY

VOLUME LXIX, NO. 13, JULY 13, 1972

# TARSKI'S THEORY OF TRUTH *

IN the early 1930s there was prevalent, among scientifically minded philosophers, the view that semantic notions such as the notions of truth and denotation were illegitimate: that they could not or should not be incorporated into a scientific conception of the world. But when Tarski's work on truth became known, all this changed. "As a result of Tarski's teaching, I no longer hesitate to speak of 'truth' and 'falsity'," wrote Popper[1]; and Popper's reaction was widely shared.[2]

A philosopher who shared Popper's reaction to Tarski's discoveries would presumably argue as follows. "What Tarski did was to define the term 'true', using in his definitions only terms that are clearly acceptable. In particular, he did not employ any undefined semantic terms in his definitions. So Tarski's work should make the term 'true' acceptable even to someone who is initially suspicious of semantic terms."

This contention has an initial plausibility, but I will argue that it is radically wrong. My contrary claim will be that Tarski succeeded in reducing the notion of truth *to certain other semantic notions*; but that he did not in any way explicate these other notions, so that his results ought to make the word 'true' acceptable only to someone who already regarded these other semantic notions as acceptable.

By claiming that Tarski merely reduced truth to other semantic notions, I don't mean to suggest that his results on truth are trivial.

* This paper grew out of a talk I gave at Princeton in the fall of 1970, where I defended T1 over T2. Donald Davidson and Gilbert Harman—and later, in private conversation, John Wallace—all came to the defense of T2, and their remarks have all been of help to me in writing the paper. I have also benefited from advice given by Michael Devitt, Paul Benacerraf, and especially David Hills.

[1] *Logic of Scientific Discovery* (New York: Basic Books, 1968), p. 274.

[2] Cf. Carnap's "Autobiography," in P. A. Schilpp, ed., *The Philosophy of Rudolf Carnap* (Lasalle, Ill.: Open Court, 1963), p. 61.

On the contrary, I think that they are extremely important, and have applications not only to mathematics but also to linguistics and to more directly "philosophical" problems about realism and objectivity. I think, however, that the real value of Tarski's discoveries for linguistics and philosophy is widely misunderstood, and I hope to eradicate the most central misunderstandings by clarifying and defending the claim that Tarski merely reduced truth to other semantic notions.

I

I believe that Tarski presented his semantic theory in a very misleading way, one which has encouraged the misinterpretations just alluded to. In this section I will present Tarski's theory as I think he should have presented it. However, I do not expect instant agreement that this new way is better than the old, and so I will use the name 'Tarski*' for a logician who gave the sort of semantic theory I will now sketch. Later in the paper I will compare Tarski*s semantics to the semantics that the real Tarski actually gave; by doing this I will cast light on the issues raised in my introductory paragraphs.

In sketching Tarski*s theory, I will focus my attention on a particular object language L. The language L that I choose will be a quantificational language with names ('$c_1$', '$c_2$', ...), one-place function symbols ('$f_1$', '$f_2$', ...), and one-place predicates ('$p_1$', '$p_2$', ...). The language of course cannot be viewed as an "uninterpreted" language, i.e., as just a bunch of strings of meaningless marks, for then there would be no truth to worry about. Instead, the language should be regarded as something that people actually speak or write; and it is because the speakers speak or write the way they do that the words of the language have the meaning they have.[3]

Initially I will follow Tarski in supposing that in L "the sense of every expression is unambiguously determined by its form,"[4] i.e., that whenever two speakers use the same name (or one speaker uses it on two occasions) they are referring to the same thing, that whenever two speakers use the same sentence either both are saying some-

[3] It is sometimes claimed that Tarski was interested in languages considered in abstraction from all speakers and writers of the language; that the languages he was dealing with are abstract entities to be specified by giving their rules. This seems incorrect: Tarski was interested in giving the semantics of languages that mathematicians had been writing for years; and only as a result of Tarski's work was it then possible for philosophers like Carnap to propose that the clauses of a Tarski-type truth definition for such languages be called rules of the languages and be used in defining the languages as abstract entities.

[4] "The Concept of Truth in Formalized Languages" (CTFL), in *Logic, Semantics, and Metamathematics* (*LSM*) (New York: Oxford, 1956), p. 166.

thing true or neither is, etc. In these circumstances it makes sense to speak of the names of the language denoting things (a name denotes whatever the users of the name refer to) and the sentences being true or false (true when speakers who use it say something true by so doing). The more general situation, in which there are expressions whose "sense" is not determined wholly by their form, will be dealt with later. (We'll see that it is one of the advantages of Tarski*s semantics that it can easily handle this more general situation.)

The syntax of L can be given by two recursive definitions: first we define the *singular terms* by saying that all names and variables are singular terms, and a function symbol followed by a singular term is a singular term; then we define the *formulas* by saying that a predicate followed by a singular term is a formula, as is the negation of a formula, the conjunction of two formulas, and the universal quantification of a formula with any variable. The *sentences*, or *closed formulas*, are then singled out in the usual way.

Now we can proceed to Tarski*s semantics. Rather than characterize truth directly, we characterize it relative to some assignment of objects to the variables, say $s_k$ to '$x_k$'. The idea is going to be to treat the variables, or at least the free variables, as sort of "temporary names" for the objects assigned to them. So we proceed by fixing a sequence $s = \langle s_1, s_2, \ldots \rangle$ of objects, to be assigned to '$x_1$', '$x_2$', ..., respectively; and we want to say what it is for a formula to be true$_s$, i.e., true relative to the assignment $s$. As a preliminary we say what it is for a term to denote$_s$ an object, i.e., to denote it relative to the assignment $s$. The denotation of '$x_k$' relative to $s$ is evidently $s_k$, for this is the object assigned to '$x_k$'. But what is the denotation relative to $s$ of '$c_k$'? Evidently what objects are assigned to the variables here is irrelevant, and the denotation$_s$ of '$c_k$' is some fixed object that users of the language refer to when they use the name '$c_k$'. Just what this object is depends on facts we have not yet been given about the use of '$c_k$'. Similarly there are facts we have not yet been given about the use of '$p_k$' and '$f_k$' which we need in order to fix the truth value of sentences containing them. For '$p_k$' the relevant facts concern the extension of the predicate—what objects the predicate *applies to*—for it is this which affects the truth value of all utterances containing '$p_k$'. For '$f_k$', the relevant facts concern what pairs of objects *fulfill* that function symbol—in the sense that the pair ⟨John Adams, John Quincy Adams⟩ and every other father-son pair fulfill the function symbol 'father of'.

With these points in mind it is now easy to give an inductive characterization of denotation$_s$:

T1 (A) 1. '$x_k$' denotes$_s$ $s_k$.
2. '$c_k$' denotes$_s$ what it denotes.
3. $\ulcorner f_k(e) \urcorner$ denotes$_s$ an object $a$ if and only if
(i) there is an object $b$ that $e$ denotes$_s$
and (ii) '$f_k$' is fulfilled by $\langle a,b \rangle$.

(Here '$e$' is a variable ranging over expressions of L.) Similarly we define 'true$_s$' for formulas—what Tarski calls satisfaction of a formula by $s$:

(B) 1. $\ulcorner p_k(e) \urcorner$ is true$_s$ if and only if
(i) there is an object $a$ that $e$ denotes$_s$
and (ii) '$p_k$' applies to $a$.
2. $\ulcorner \sim e \urcorner$ is true$_s$ if and only if $e$ is not true$_s$.
3. $\ulcorner e_1 \wedge e_2 \urcorner$ is true$_s$ if and only if $e_1$ is true$_s$ and so is $e_2$.
4. $\ulcorner \forall x_k(e) \urcorner$ is true$_s$ if and only if for each sequence $s^*$ that differs from $s$ at the $k$th place at most, $e$ is true$_{s^*}$.

This completes the characterization of truth relative to an assignment of objects to the variables. In the case of sentences it is easily seen that we get the same results whatever such assignment we pick; we can say

(C) A sentence is true if and only if its is true$_s$ for some (or all) $s$.

This completes my elaboration of Tarski*s "truth definition" T1 for L—or his *truth characterization* (TC), as I prefer to call it. What is its philosophical significance? The obvious answer, and the correct one, I think, is that the TC reduces one semantic notion to three others. It explains what it is for a sentence to be true in terms of certain semantic features of the primitive components of the sentence: in terms of what it is for a name to denote something, what it is for a predicate to apply to something, and what it is for a function symbol to be fulfilled by some pair of things. It is convenient to introduce the expression 'primitively denotes' as follows: every name *primitively denotes* what it denotes; every predicate and every function symbol *primitively denotes* what it applies to or is fulfilled by; and no complex expression primitively denotes anything. In this terminology, what T1 does is to explain truth in terms of primitive denotation. Similarly we can explain denotation for arbitrary closed singular terms [such as '$f_1(c_1)$'] in terms of primitive denotation, i.e., in terms of the semantic features of the names and function symbols from which the complex singular term is composed—we

merely say that a closed singular term denotes an object $a$ if it denotes$_s$ $a$ for some (or all) $s$, where denotation$_s$ is defined as before. We see then that *Tarski*s semantics explains the semantic properties of complex expressions* (e.g., truth value for sentences, denotation for complex singular terms) *in terms of semantic properties of their primitive components.*

To explain truth in terms of primitive denotation is, I think, an important task. It certainly doesn't answer *every* question that anyone would ever want answered about truth, but for many purposes it is precisely what we need. For instance, in model theory we are interested in such questions as: given a set $\Gamma$ of sentences, is there any way to choose the denotations of the primitives of the language so that every sentence of $\Gamma$ will come out true given the usual semantics for the logical connectives?[5] For questions such as this, what we need to know is how the truth value of a whole sentence depends on the denotations of its primitive nonlogical parts, and that is precisely what T1 tells us. So *at least for model-theoretic purposes*, Tarski*s TC is precisely the kind of explication of truth we need.

I want now to return to a point I mentioned earlier, about Tarski's restriction to languages in which "the sense of every expression is unambiguously determined by its form." Natural languages are full of expressions that do not meet this requirement. For instance, different tokens of 'John takes grass' can differ in "sense" —e.g., one token may be uttered in saying that John Smith smokes marijuana, and another may be uttered in saying that John Jones steals lawn material, and these differences may give rise to differences of truth value in the tokens. (I say that a complete[6] token of a sentence is true if the person who spoke or wrote that token said something true by so doing; I also say that a name token denotes an object if the person who spoke or wrote the token referred to the object by so doing.) The prevalence of such examples in natural languages raises the question of whether Tarski's type of semantic theory is applicable to languages in which the sense is *not* determined by the form; for if the answer is no, then Davidson's very worth-

[5] Actually in model theory we are interested in allowing a slightly unusual semantics for the quantifiers: we are willing to allow that the quantifier not range over everything. We could build this generalization into our truth definition, by stipulating that in addition to the denotations of the nonlogical symbols we specify a universe U, and then reformulating clause (B)4 by requiring that the $k$th member of $s^*$ belong to U. If we did this, then it would be the range of the quantifiers as well as the denotations of the nonlogical primitives that we would have explained truth in terms of.

[6] An *incomplete* sentence token is a sentence token which [like the occurrence of '$2 + 2 = 4$' inside '$\sim(2 + 2 = 4)$'] is part of a larger sentence token.

while project[7] of giving truth characterizations for natural languages seems doomed from the start.

It seems clear that if we stick to the kind of TC that Tarski actually gave (see next section), there is no remotely palatable way of extending TC's to sentences like 'John takes grass'. But if we use TC's like T1 there is no difficulty at all. The only point about languages containing 'John' or 'grass' or 'I' or 'you' is that for such languages 'true', 'denotes', and other semantic terms make no clear sense as applied to expression types; they make sense only as applied to tokens. For this reason we have to interpret clause (B)2 of T1 as meaning

> A token of $\ulcorner\sim e\urcorner$ is true$_s$ if and only if the token of $e$ that it contains is not true$_s$.

and similarly for the other clauses. Once we interpret our TC in this way in terms of tokens, i.e., individual occasions of utterance, that TC works perfectly: someone who utters 'John is sick' (or 'I am sick') says something true if and only if his token of 'sick' applies to the person he refers to by 'John' (or by 'I'); and the fact that other speakers (or this speaker on other occasions) sometimes refer to different things when they use 'John' (or 'I') is beside the point.

This analysis leaves entirely out of account the ways in which 'I' and 'John' differ: it leaves out of account, for instance, the fact that a token of 'I' always denotes the speaker who produced it. But that is no objection to the analysis, for the analysis purports merely to explain truth in terms of primitive denotation; it does not purport to say anything about primitive denotation, and the differences between 'I' and 'John' (or their analogues in a language like L) are purely differences of how they denote. (The word 'I' denotes according to the simple rule mentioned two sentences back; 'John' denotes according to much more complex rules that I have no idea how to formulate.)

Of course, the fact that a theory of denotation for a word like 'I' is so simple and obvious, makes it possible to alter the TC so that the theory of denotation for such a word is built into the TC itself—such a course is adopted, for instance, by Davidson at the end of "Truth and Meaning." I myself prefer to preserve the analogies of the word 'I' to words that function less systematically, e.g., 'we', 'she', and 'John'. How one treats 'I' is more or less a matter of taste; but the less systematic words I've just mentioned cannot be handled in the way that Davidson handles 'I', and the only reasonable way I can

[7] "Truth and Meaning," *Synthese*, XVII, 3 (September, 1967): 304–323, pp. 314/5.

see to handle them is the way I have suggested: use a truth characterization like T1 (except stated in terms of tokens rather than types), and leave it to a separate theory of primitive denotation to explain the relevant differences between tokens of 'John' that denote John Adams and tokens of 'John' that denote John Lennon, and between tokens of 'bank' that apply to things along rivers and tokens of 'bank' that apply to the Chase Manhattan.[8]

There are other advantages to T1 besides its ability to handle ambiguous sentences, i.e., sentences for which the sense is not determined by the form. For instance, Tarski required that the vocabulary of the language be fixed once and for all; but if we decide to give truth characterizations of type T1, this is unnecessary: all that is required is that the general structure of the language be fixed, e.g., that the semantic categories[9] (name, one-place predicate, etc.) be held constant. In other words, if a language already contained proper names, the invention of a new name to baptize an object will not invalidate the old TC; though introduction of a name into a hitherto nameless language will.

To show this, we have merely to reformulate the given TC so that it does not rely on the actual vocabulary that the language contains at a given time, but works also for sentences containing new names, one-place predicates, etc., that speakers of the language might later introduce. To do this is trivial: we define denotation$_s$ by

1. The $k$th variable denotes$_s$ $s_k$.
2. If $e_1$ is a name, it denotes$_s$ what it denotes.
3. If $e_1$ is a singular term and $e_2$ is a function symbol, then $\ulcorner e_2(e_1)\urcorner$ denotes$_s$ $a$ if and only if
   (i) as before,
   and (ii) $e_2$ is fulfilled by $\langle a,b\rangle$.

and we can generalize the definition of truth$_s$ in a similar manner.[10] This shows that, in giving a TC, there is no need to utilize the particular vocabulary used at one temporal stage of a language, for we

[8] Note that the claims I've been making are intended to apply only to cases where different tokens have different semantic features; they are not intended to apply to cases of indeterminacy, i.e., to cases where a particular name token or predicate token has no determinate denotation or extension. To deal with indeterminacy requires more complex devices than I employ in this paper.

[9] The notion of a semantic category is Tarski's: cf. CTFL, p. 215.

[10] To do so in the obvious way requires that we introduce semantic categories of negation symbol, conjunction symbol, and universal-quantification symbol; though by utilizing some ideas of Frege it could be shown that there is really no need of a separate semantic category for each logical operator. The use of semantic categories in the generalized truth characterization raises important problems which I have had to suppress for lack of space in this paper.

can instead give a more general TC which can be incorporated into a diachronic theory of the language (and can also be applied directly to other languages of a similar structure). *If*, that is, we accept the modification of Tarski proposed in this section.

II

The kind of truth characterization advocated in the previous section differs from the kind of TC Tarski offered in one important respect. Tarski stated the policy "I shall not make use of any semantical concept if I am not able previously to reduce it to other concepts" (CTFL 152/3), and this policy is flagrantly violated by T1: T1 utilizes unreduced notions of proper names denoting things, predicates applying to things, and function symbols being fulfilled by things.

Tarski's truth characterizations, unlike T1, accorded with his stated policy: they did not contain any semantic terms like 'applies to' or 'denotes'. How did Tarski achieve this result? Very simply: first, he translated every name, predicate, and function symbol of L into English; then he utilized these translations in order to reformulate clauses 2 and 3(ii) of part (A) of the definition and clause 1(ii) of part (B). For simplicity, let's use '$\bar{c}_1$', '$\bar{c}_2$', etc. as abbreviations for the English expressions that are the translations of the words '$c_1$', '$c_2$',... of L: e.g.: if L is simplified German and '$c_1$' is 'Deutschland', then '$\bar{c}_1$' is an abbreviation for 'Germany'. Similarly, let '$\bar{f}_1$' abbreviate the translation into English of the word '$f_1$' of L, and let '$\bar{p}_1$' abbreviate the translation of '$p_1$' into English. Then Tarski's reformulated truth definition will read as follows:

T2 (A) 1. as before
2. '$c_k$' denotes$_s$ $\bar{c}_k$
3. $\ulcorner f_k(e) \urcorner$ denotes$_s$ $a$ if and only if
(i) as before
(ii) $a$ is $\bar{f}_k(b)$

(B) 1. $\ulcorner p_k(e) \urcorner$ is true$_s$ if and only if
(i) as before
(ii) $\bar{p}_k(a)$

2–4. as before

(C) as before

What T2 is like depends of course on the precise character of the translations of the primitives that are utilized. For instance, if we translate '$c_1$' as 'the denotation of '$c_1$' ', translate '$p_1$' as 'is something that '$p_1$' applies to', etc., then T2 becomes identical with T1. This of course is *not* what Tarski intended. What Tarski intended is that T2 not contain unexplicated semantic terms, and if we are to get

this result we must not employ any semantic terms in our translations.[11]

But other restrictions on translations are also necessary: if we were to translate 'Deutschland' as 'Bertrand Russell', a truth characterization T2 that was based on this translation would grossly misrepresent L. In order to state the matter more generally, I introduce the term 'coreferential': two singular terms are coreferential if they denote the same thing; two predicative expressions are coreferential if they have the same extension, i.e., if they apply to the same things; and two functional expressions are coreferential if they are fulfilled by the same pairs. It is then easily seen that any departure from coreferentiality in translation will bring errors into T2. For instance, suppose we translate the foreign predicate 'glub' as 'yellow', and suppose 'glub' and yellow are not *precisely* coreferential; then clause $(B)_1$ will say falsely that 'glub$(x)$' is true of just those objects which are yellow.

Let us say, then, that

(1) An adequate translation of a primitive $e_1$ of $L$ into English is an expression $e_2$ of English such that
  (i) $e_1$ and $e_2$ are coreferential, and
  (ii) $e_2$ contains no semantic terms.

This notion of an adequate translation is of course a semantic notion that Tarski did not reduce to nonsemantic terms. But that is no objection to his characterization T2 (at least, it isn't obviously an objection), for the notion of an adequate translation is never built into the truth characterization and is not, properly speaking, part of a theory of truth. On Tarski's view we need to adequately translate the object language into the metalanguage in order to give an adequate theory of truth for the object language; this means that the notion of an adequate translation is employed in the methodology of giving truth theories, but it is not employed in the truth theories themselves.

In what follows I shall assume that the language L with which we are dealing is so related to English that all its primitives *can* be adequately translated into English, according to the standards of adequacy set forth in (1). (This is another restriction that we avoid if we give TC's of the type T1; quite a significant restriction, I think.) If we then suppose that the translation given ('$\bar{c}_1$' for '$c_1$', etc.) is one of the adequate translations, then T2, like T1, is a correct recursive characterization of truth for the language L. There is, of course, a

[11] For simplicity, I have assumed that L itself contains no semantic terms.

simple procedure for transforming recursive characterizations such as these into explicit characterizations. To carry the procedure through in these cases would be pretty complicated, but it could be done; so we could regard T1 (or T2) as implicitly specifying a metalinguistic formula '$A_1(e)$' (or '$A_2(e)$'), and saying that an utterance $e$ of L is true if and only if $A_1(e)$ (or $A_2(e)$). If we regard T1 and T2 as written in this form, then the key difference between them is that '$A_1(e)$' *contains semantic terms and* '$A_2(e)$' *does not.* The question then arises: is the fact that '$A_2(e)$' does not contain semantic terms an advantage of T2 over T1? If so, then *why* is it an advantage?

In order to discuss the possible advantages of T2 over T1, I think we have to go beyond mathematical considerations and focus instead on linguistic and other "philosophical" matters. It is not enough to say that T2 *defines* truth without utilizing semantic terms, whereas T1 defines it only in other semantic terms: this is not enough until we say something more about the purpose of definition. If the purpose of giving a "definition" of truth is to enable you to do model theory, then the elimination of semantic terms from T1 gives no advantage. For what purpose do we want definitions for which the elimination of semantic terms is useful?

One purpose to which definitions are sometimes put is in explaining the meaning of a word. This of course is very vague, but I think it is clear enough to enable us to recognize that neither T1 nor T2 has very much to do with explaining the meaning of the word 'true'. This is especially obvious for T2: a T2-type truth definition works for a single language only, and so if it "explains the meaning of" the word 'true' as applied to that language, then for *any* two languages $L_1$ and $L_2$, the word 'true' means something different when applied to utterances of $L_1$ than it means when applied to utterances of $L_2$! I make this point not in criticism of T2, but in criticism of the idea that the significance of T2 can be explained by saying that it "gives the meaning of" the word 'true'.

We still need to know what purpose a truth characterization like T1 or T2 could serve that would give someone reason to think that a TC without unexplicated semantic terms would be better than a TC with unexplicated semantic terms. Tarski hints at such a purpose in one place in his writings, where he is discussing the importance of being able to define the word 'true', as opposed to merely introducing axioms to establish the basic properties of truth. If a definition of semantic notions such as truth could not be given, Tarski writes,

> . . . it would then be difficult to bring [semantics] into harmony with the postulates of the unity of science and of physicalism (since

the concepts of semantics would be neither logical nor physical concepts).[12]

This remark seems to me to be of utmost importance in evaluating the philosophical significance of Tarski's work, and so I will now say something about the general philosophical issues it raises. When this is done we will be in a better position to understand Tarski's choice of T2 over T1.

III

In the early 1930s many philosophers believed that the notion of truth could not be incorporated into a scientific conception of the world. I think that the main rationale for this view is hinted at in the remark of Tarski's that I quoted at the end of the last section, and what I want to do now is to elaborate a bit on Tarski's hint.

In the remark I have quoted, Tarski put a heavy stress on the doctrine of physicalism: the doctrine that chemical facts, biological facts, psychological facts, and semantical facts, are all explicable (in principle) in terms of physical facts. The doctrine of physicalism functions as a high-level empirical hypothesis, a hypothesis that no small number of experiments can force us to give up. It functions, in other words, in much the same way as the doctrine of mechanism (that all facts are explicable in terms of *mechanical* facts) once functioned: this latter doctrine has now been universally rejected, but it was given up only by the development of a well-accepted theory (Maxwell's) which described phenomena (electromagnetic radiation and the electromagnetic field) that were very difficult to account for mechanically, and by amassing a great deal of experiment and theory that together made it quite conclusive that mechanical explanations of these phenomena (e.g., by positing "the ether") would never get off the ground. Mechanism has been empirically refuted; its heir is physicalism, which allows as "basic" not only facts about mechanics, but facts about other branches of physics as well.[13] I believe that physicists a hundred years ago were justified in accepting mechanism, and that, similarly, physicalism should be accepted until we have convincing evidence that there is a realm of phenomena it leaves out of account. Even if there *does* turn out to be such a realm of phenomena, the only way we'll ever come to know

[12] "The Establishment of Scientific Semantics" (ESS) in *LSM*, p. 406.

[13] This, of course, is very vague, but most attempts to explicate the doctrine of physicalism more precisely result in doctrines that are very hard to take seriously [e.g., the doctrine that for every acceptable predicate '$P(x)$' there is a formula '$B(x)$' containing only terminology from physics, such that '$\forall x(P(x) \equiv B(x))$' is true]. Physicalism should be understood as the doctrine (however precisely it is to be characterized) that guides science in the way I describe.

that there is, is by repeated efforts and repeated failures to explain these phenomena in physical terms.

That's my view, anyway, but there are philosophers who think that it is in order to reject physicalism now. One way of rejecting physicalism is called "vitalism": it is the view that there are irreducibly biological facts, i.e., biological facts that aren't explicable in nonbiological terms (and hence, not in physical terms). Physicalism and vitalism are incompatible, and it is because of this incompatibility that the doctrine of physicalism has the methodological importance it has for biology. Suppose, for instance, that a certain woman has two sons, one hemophilic and one not. Then, according to standard genetic accounts of hemophilia, the ovum from which one of these sons was produced must have contained a gene for hemophilia, and the ovum from which the other son was produced must not have contained such a gene. But now the doctrine of physicalism tells us that there must have been a *physical* difference between the two ova that explains why the first son had hemophilia and the second one didn't, if the standard genetic account is to be accepted. We should not rest content with a special biological predicate 'has-a-hemophilic-gene'—rather, we should look for nonbiological facts (chemical facts; and ultimately, physical facts) that underlie the correct application of this predicate. That at least is what the principle of physicalism tells us, and it can hardly be doubted that this principle has motivated a great deal of very profitable research into the chemical foundations of genetics.

So much for vitalism; now let us turn to other irreducibility doctrines that are opposed to physicalism. One such irreducibility doctrine is Cartesianism: it is the doctrine that there are irreducibly mental facts. Another irreducibility doctrine has received much less attention than either vitalism or Cartesianism, but it is central to our present concerns: this doctrine, which might be called "semanticalism," is the doctrine that there are irreducibly semantic facts. The semanticalist claims, in other words, that semantic phenomena (such as the fact that 'Schnee' refers to snow) must be accepted as primitive, in precisely the way that electromagnetic phenomena are accepted as primitive (by those who accept Maxwell's equations and reject the ether); and in precisely the way that biological phenomena and mental phenomena are accepted as primitive by vitalists and Cartesians. Semanticalism, like Cartesianism and vitalism, posits nonphysical primitives, and as a physicalist I believe that all three doctrines must be rejected.

There are two general sorts of strategy that can be taken in rejecting semanticalism, or Cartesianism, or vitalism. One strategy,

illustrated two paragraphs back in discussing vitalism, is to try to explicate the terms of a biological theory in nonbiological terms. But there is another possible strategy, which is to argue that the biological terms are illegitimate. The second strategy seems reasonable to adopt in dealing with the following predicate of (reincarnationist) biology: '$x$ has the same soul as $y$'. A physicalist would never try to find physical or chemical facts that underlie reincarnation; rather, he would reject reincarnation as a myth.

Since biological theory is as well developed as it is, we usually have a pretty good idea which biological terms require explication and which require elimination. When we turn to psychology and semantics, however, it is often not so obvious which strategy is the more promising. Thus in semantics, physicalists agree that all *legitimate* semantic terms must be explicable nonsemantically—they think in other words that there are no irreducibly semantic facts—but they disagree as to which semantic terms are legitimate. That disagreement has become fairly clear in recent years in the theory of meaning, with the work of Quine: the disagreement is between those physicalists who would look for a nonsemantic basis for terms in the theory of meaning, and those who would follow Quine in simply throwing out those terms. Our concern, however, is not with the theory of meaning, but with the theory of reference, and here the disagreement has been less clear, since there haven't been many physicalists who openly advocate getting rid of terms like 'true' and 'denotes'. There were such physicalists in the early 1930s; part of the importance of Tarski's work was to persuade them that they were on the wrong track, to persuade them that we should explicate notions in the theory of reference nonsemantically rather than simply get rid of them.

The view that we should just stop using semantic terms (here and in the rest of this paper, I mean terms in the theory of reference, such as 'true' and 'denotes' and 'applies to') draws its plausibility from the apparent difficulty of explicating these terms nonsemantically. People utter the sounds 'Electrons have rest mass but photons don't', or 'Schnee ist weiss und Gras ist grün', and we apply the word 'true' to their utterances. We don't want to say that it is a primitive and inexplicable fact about these utterances that they are true, a fact that cannot be explicated in nonsemantic terms; this is as unattractive to a physicalist as supposing that it is a primitive and inexplicable fact about an organism at a certain time that it is in pain. But how could we ever explicate in nonsemantic terms the alleged fact that these utterances are true? *Part* of the explication of the truth of 'Schnee ist weiss und Gras ist grün', presumably, would be

that snow is white and grass is green. But this would only be part of the explanation, for still missing is the connection between snow being white and grass being green on the one hand, and the German utterance being true on the other hand. It is this connection that seems so difficult to explicate in a way that would satisfy a physicalist, i.e., in a way that does not involve the use of semantic terms.

If, in face of these difficulties, we were ever to conclude that it was *impossible* to explicate the notions of truth and denotation in nonsemantic terms, we would have either to give up these semantic terms or else to reject physicalism. It seems to me that that is essentially what Tarski is saying in the quotation at the end of the last section, and I have tried to make it plausible by sketching analogies to areas other than semantics. Tarski's view, however, was that, for certain languages at least, semantic terms *are* explicable nonsemantically, and that truth definitions like T2 provide the required explication. It is understandable that as far as *philosophical* purposes go Tarski should think that T1 leaves something to be desired: after all, it merely explicates truth in terms of other semantic concepts; but what good does that do if those other concepts can't be explicated nonsemantically? T2, then, has a strong prima facie advantage over T1. In the next section I will show that it is not a genuine advantage.

## IV

The apparent advantage of T2 over T1, I have stressed, is that it appears to reduce truth to nonsemantic terms; and I *think* this is why Tarski wanted to give a truth definition like T2 rather than like T1. This interpretation makes sense of Tarski's remark about physicalism, and it also explains why someone who was certainly not interested in "meaning analysis" as that is usually conceived would have wanted to give "definitions" of truth and would emphasize that, in these "definitions," "I will not make use of any semantical concept if I am not able previously to reduce it to other concepts." In any case, the problem of reducing truth is a very important problem, one which T1 and T2 provide a partial solution to, and one which T2 *might* be thought to provide a full solution to; and it is not at all clear what *other* interesting problems T2 could be though to solve better than T1.

In Tarski's own exposition of his theory of truth, Tarski put very little stress on the problem of reduction or on any other problem with a clear philosophical or mathematical motivation; instead, he set up a formal criterion of adequacy for theories of truth without any serious discussion of whether or why this formal criterion is

reasonable. Roughly, the criterion was this:[14]

(M) Any condition of the form

(2) $(\forall e)[e \text{ is true} \equiv B(e)]$

should be accepted as an adequate definition of truth if and only if it is correct and '$B(e)$' is a well-formed formula containing no semantic terms. (The quantifiers are to be taken as ranging over expressions of one particular language only.)

The "only if" part of condition M is not something I will contest.

[14] Tarski actually gives a different formulation, the famous Convention T, evidently because he does not think that the word 'correct' ought to be employed in stating a criterion of adequacy. First of all Tarski writes

> . . . we shall accept as valid every sentence of the form
>
> [T] the sentence $x$ is true if and only $p$
>
> where '$p$' is to be replaced by any sentence of the language under investigation and '$x$' by any individual name of that sentence provided this name occurs in the metalanguage (ESS 404).

Is Tarski's policy of accepting these sentences as "valid" (i.e., true) legitimate? It seems to me that it is, in a certain special case. The special case is where

I. The object language is a proper part of the metalanguage (here, English).

II. The object language contains no paradoxical or ambiguous or truth-value-less sentences.

In this special case—and it was the case that Tarski was primarily concerned with —I think it will be generally agreed that all instances of Schema T hold. From this, together with the fact that only grammatical sentences are true, we can argue that, if a necessary and sufficient condition of form (2) has the following consequences:

(a) Every instance of Schema T

(b) The sentence '$(\forall x)(x \text{ is true} \supset S(x))$', where '$S(x)$' formulates (correct) conditions for an utterance of L to be a sentence

then that necessary and sufficient condition is correct. Let's say that a "truth definition" for L (a necessary and sufficient condition of truth in L) *satisfies Convention T* if it has all the consequences listed under (a) and (b). Then, restating: when L is a language for which I and II hold, then any truth definition satisfying Convention T is correct; and since only quite uncontroversial assumptions about truth are used in getting this result, anyone will admit to the correctness of a truth characterization satisfying Convention T. If we use the term 'formally correct definition' for a sentence of form (2) in which '$B(e)$' contains no semantic terms, this means that a formally correct definition that satisfies Convention T is bound to satisfy Condition M (when the language L satisfies I and II). As far as I can see, this is the only motivation for Convention T.

Tarski sometimes states a more general form of Convention T, which applies to languages that do not meet restriction I: it is what results when one allows as instances of Schema T the results of replacing '$p$' by a *correct translation* of the sentence that the name substituted for '$x$' denotes (in some sense of 'correct translation' in which correctness requires preservation of truth value). But then the advantage of the ungeneralized form of Convention T (viz., that anything satisfying it wears its correctness on its face, or more accurately, on the faces of its logical consequences) is lost.

It rules out the possibility of T1 *by itself* being an adequate truth definition; and it is right to do so, if the task of a truth definition is to reduce truth to nonsemantic terms, for T1 provides only a *partial* reduction. (To complete the reduction we need to reduce primitive denotation to nonsemantic terms.) T2, on the other hand, meets condition M; so either T2 is superior to T1 as a reduction, or else condition M is too weak and the "if" part of it must be rejected. My own diagnosis is the latter, but the other possibility seems initially reasonable. After all, how could condition M be strengthened? We might try requiring that '*B*(*e*)' be not only *extensionally* equivalent to '*e* is true', but *intensionally* equivalent to it; but this clearly won't do, for even if we grant that there is an intelligible notion of intensional equivalence, our concern is not with analyzing the meaning of the word 'true' but with performing a reduction. A clear and useful standard of equivalence that is stronger than extensional equivalence but not so strong as to rule out acceptable reductions is unknown at the present time, so I know no way to improve on condition M. My view is that we have a rough but useful concept of reduction which we are unable to formulate precisely; but I must admit that the alternative view, that extensional equivalence is adequate, has an initial appeal.

A closer look, however, will reveal quite conclusively that extensional equivalence is not a sufficient standard of reduction. This can be seen by looking at the concept of valence. The valence of a chemical element is an integer that is associated with that element, which represents the sort of chemical combinations that the element will enter into. What I mean by the last phrase is that it is possible—roughly, at least—to characterize which elements will combine with which others, and in what proportions they will combine, merely in terms of their valences. Because of this fact, the concept of valence is a physically important concept, and so if physicalism is correct it ought to be possible to explicate this concept in physical terms—e.g., it ought to be possible to find structural properties of the atoms of each element that determine what the valence of that element will be. Early in the twentieth century (long after the notion of valence had proved its value in enabling chemists to predict what chemical combinations there would be) this reduction of the concept of valence to the physical properites of atoms was established; the notion of valence was thus shown to be a physicalistically acceptable notion.

Now, it would have been easy for a chemist, late in the last century, to have given a "valence definition" of the following form:

(3) $(\forall E)(\forall n)(E$ has valence $n \equiv E$ is potassium and $n$ is $+1$, or . . . or $E$ is sulphur and $n$ is $-2)$

where in the blanks go a list of similar clauses, one for each element. But, though this is an extensionally correct definition of valence, it would not have been an acceptable reduction; and had it turned out that nothing else was possible—had all efforts to explain valence in terms of the structural properties of atoms proved futile—scientists would have eventually had to decide either (a) to give up valence theory, or else (b) to replace the hypothesis of physicalism by another hypothesis (chemicalism?). It is part of scientific methodology to resist doing (b); and I also think it is part of scientific methodology to resist doing (a) as long as the notion of valence is serving the purposes for which it was designed (i.e., as long as it is proving useful in helping us characterize chemical compounds in terms of their valences). But the methodology is not to resist (a) and (b) by giving lists like (3); the methodology is to look for a real reduction. This is a methodology that has proved extremely fruitful in science, and I think we'd be crazy to give it up in linguistics. *And I think we are giving up this fruitful methodology, unless we realize that we need to add theories of primitive reference to* T1 *or* T2 *if we are to establish the notion of truth as a physicalistically acceptable notion.*

I certainly haven't yet given much argument for this last claim. I *have* argued that the standard of extensional equivalence doesn't guarantee an acceptable reduction; but T2 is obviously not trivial to the extent that (3) is. What *is* true, however, is roughly that T2 minus T1 is as trivial as (3) is. One way in which this last claim can be made more precise is by remembering that really we often apply the term 'valence' not only to elements, but also to configurations of elements (at least to stable configurations that are not compounds, i.e., to radicals). Thus, if we abstract from certain physical limitations on the size of possible configurations of elements (as, in linguistics, we usually abstract from the limitations that memory, etc., impose on the lengths of possible utterances), there is an infinite number of entities to which the term 'valence' is applied. But it is an important fact about valence that the valence of a configuration of elements is determined from the valences of the elements that make it up, and from the way they're put together. Because of this, we might try to give a recursive characterization of valence. First of all, we would try to characterize all the different *structures* that configurations of elements can have (much as we try to characterize all the different grammatical structures before we give a truth definition like T1 or T2). We would then try to find rules that would

enable us to determine what the valence of a complicated configuration would be, given the valences of certain less complicated configurations that make it up and the way they're put together. If we had enough such rules, we could determine the valence of a given configuration given only its structure and the valences of the elements that make it up. And if we like, we can transform our recursive characterization of valence into an explicit characterization, getting

$$\text{V1}\quad (\forall c)(\forall n)\ (c \text{ has valence } n \equiv B(c,n))$$

The formula '$B(c,n)$' here employed will still contain the term 'valence', but it will contain that term only as applied to elements, not as applied to configurations. Thus our "valence definition" V1 would characterize the valence of the complex *in terms of the valences of the simple.*

It would now be possible to eliminate the term 'valence' from '$B(c,n)$', in either of two ways. One way would be to employ a genuine reduction of the notion of valence for elements to the structural properties of atoms. The other way would be to employ the pseudo-reduction (3). It is clear that we could use (3) to give a trivial reformulation V2 of V1, which would have precisely the "advantages" as a reduction that T2 has over T1. (V2, incidentally, would also have one of the disadvantages over V1 that T2 has over T1: V1 does not need to be overhauled when you discover or synthesize new elements, whereas V2 does.)

That is a sketch of one way that the remark I made two paragraphs back about "T2 minus T1" could be made more precise. But it is somewhat more fruitful to develop the point slightly differently: doing this will enable me to make clearer that there is unlikely to be *any* purpose that T2 serves better than T1 (not merely that T2 is no better at reduction).

To get this result I'll go back to my original use of the term 'valence', where it applies to elements only and not to configurations. And what I will do is compare (3) not to Tarski's theory of *truth*, but to Tarski's theory of *denotation* for names; the effect of this on his theory of truth will then be considered. Tarski states his theory of denotation for names in a footnote, as follows:

> To say that the name $x$ denotes a given object $a$ is the same as to stipulate that the object $a$ . . . satisfies a sentential function of a particular type. In colloquial language it would be a function which consists of three parts in the following order: a variable, the word 'is' and the given name $x$ (CTFL 194).

This is actually only part of the theory, the part that defines denotation in terms of satisfaction; to see what the theory looks like when all semantic terms are eliminated, we must see how satisfaction is defined. The definition is given by the (A) and (B) clauses of T2, for, as I've remarked, 'satisfaction' is Tarski's name for what I've called "truth$_s$". What Tarski's definition of satisfaction tells us is this: for any name $N$, an object $a$ satisfies the sentential function $\ulcorner x_1$ is $N\urcorner$ if and only if $a$ is France and $N$ is 'France' or . . . or $a$ is Germany and $N$ is 'Germany'. Combining this definition of satisfaction (for sentential functions of form $\ulcorner x_1$ is $N\urcorner$) with the earlier account of denotation in terms of satisfaction, we get:

(DE): To say that the name $N$ denotes a given object $a$ is the same as to stipulate that either $a$ is France and $N$ is 'France', or . . . or $a$ is Germany and $N$ is 'Germany'.

This is Tarski's account of denotation for English proper names. For foreign proper names, the definition of denotation in terms of satisfaction needs no modification (except that the 'is' must be replaced by a name of a foreign word, say 'ist' for German). Combining this with the definition (again given by T2) of satisfaction for foreign sentential functions like $\ulcorner x_1$ ist $N\urcorner$, we get:

(DG): To say that the name $N$ denotes a given object $a$ is the same as to stipulate that either $a$ is France and $N$ is 'Frankreich', or . . ., or $a$ is Germany and $N$ is 'Deutschland'.

DE and DG have not received much attention in commentaries on Tarski, but in fact they play a key role in his semantic theory; and it was no aberration on Tarski's part that he offered them as theories of denotation for English and German names, for *they satisfy criteria of adequacy exactly analogous to the criteria of adequacy that Tarski accepted for theories of truth.*[15] Nevertheless, it seems clear that DE and DG do not really reduce truth to nonsemantic terms, any more than (3) reduces valence to nonchemical terms. What would a real explication of denotation in nonsemantic terms be like? The "classical" answer to this question (Russell's) is that a name

[15] A sentence of the form '$(\forall N)(\forall x)[N$ denotes $x \equiv B(N, x)]$' *satisfies convention* D if it has as consequences every instance of the schema '$y$ denotes $z$', in which '$y$' is to be replaced by a quotation-mark name for a name $N$, and '$z$' is to be replaced by (an adequate translation of $N$ into English, i.e.) a singular term of English that contains no semantic terms and that denotes the same thing that $N$ denotes. Clearly DE and DG are not only extensionally correct, they also satisy Convention D. Presumably philosophers who are especially impressed with Convention T will be equally impressed with this fact, but they owe us a reason why satisfying Convention D is of any interest.

like 'Cicero' is "analytically linked" to a certain description (such as 'the denouncer of Catiline'); so to explain how the name 'Cicero' denotes what it does you merely have to explain

> (i) the process by which it is linked to the description (presumably you bring in facts about how it was learned by its user, or facts about what is going on in the user's brain at the time of the using)
> and (ii) how the description refers to what it does

Because of (ii), of course, the project threatens circularity: the project is to explain how names refer in terms of how descriptions refer; but the natural way to explain how descriptions refer is in terms of how they're built up from their significant parts,[16] and how those significant parts refer (or apply, or are fulfilled), and those significant parts will usually include names. But Russell recognized this threat of circularity, and carefully avoided it: he assumed that the primitives of the language were to be partially ordered by a relation of "basicness," and that each name except a most basic ("logically proper") name was to be analytically linked to a formula containing only primitives more basic than it. The most basic primitives were to be linked to the world without the intervention of other words, by the relation of acquaintance.

This classical view of how names (and other primitives) latch onto their denotations is extremely implausible in many ways (e.g., it says you can refer only to things that are definable from "logically proper" primitives; it requires that there be certain statements, such as 'If Cicero existed then Cicero denounced Catiline', which are analytic in the sense that they are guaranteed by linguistic rules and are immune to revision by future discoveries). I conjecture that it is because of the difficulties with this classical theory, which was the only theory available at the time that Tarski wrote, that Tarski's pseudo-theories DE and DG seemed reasonable—they weren't exciting, but if you wanted something exciting you got logically proper names. The diagnosis that any attempt to explain the relation between words and the things they are about must inevitably lead to either a wildly implausible theory (like Russell's) or a trivial theory (like Tarski's) seems to be widely accepted still; but I think that the diagnosis has become less plausible in recent years through the

[16] For example, by extending our definition of denotation$_s$ to descriptions by:

$\ulcorner \imath x_k(e) \urcorner$ denotes$_s$ $a$ if and only if [for each sequence $s^*$ which differs from $s$ at the $k$th place at most, $e$ is true$_{s^*}$ if and only if the $k$th member of $s^*$ is $a$].

and then defining denotation in terms of denotation$_s$ by stipulating that a closed term denotes an object if and only if it denotes$_s$ that object for some (or all) $s$.

development of *causal* theories of denotation by Saul Kripke[17] and others. According to such theories, the facts that 'Cicero' denotes Cicero and that 'muon' applies to muons are to be explained in terms of certain kinds of causal networks between Cicero (muons) and our uses of 'Cicero' ('muon'): causal connections both of a social sort (the passing of the word 'Cicero' down to us from the original users of the name, or the passing of the word 'muon' to laymen from physicists) and of other sorts (the evidential causal connections that gave the original users of the name "access" to Cicero and gave physicists "access" to muons). I don't think that Kripke or anyone else thinks that *purely* causal theories of primitive denotation can be developed (even for proper names of past physical objects and for natural-kind predicates); this however should not blind us to the fact that he has suggested a kind of factor involved in denotation that gives new hope to the idea of explaining the connection between language and the things it is about. It seems to me that the possibility of *some such* theory of denotation (to be deliberately very vague) is essential to the joint acceptability of physicalism and the semantic term 'denotes', and that denotation definitions like DE and DG merely obscure the need for this.

It might be objected that the purpose of DE and DG was not reduction; but what was their purpose? One answer might be that (DE) and (DG) enable us to eliminate the word 'denote' whenever it occurs. ("To explain is to show how to eliminate.") For instance,

(4) No German name now in use denotes something that does not yet exist.

would become

(4′) For any name *N* now in use, if *N* is 'Frankreich' then France already exists, and . . ., and if *N* is 'Deutschland' then Germany already exists.

provided that (DG) is a correct and complete list of the denotations of all those German proper names that have denotations. It seems reasonably clear that we could specify a detailed procedure for transforming sentences like (4) into materially equivalent sentences like (4′). A similar claim could be made for the "valence definition" (3). Such a valence definition makes it possible to eliminate the word

[17] Some of Kripke's work on names will be published shortly in Davidson and Harman, eds., *Semantics of Natural Language* (Dordrecht: Reidel, 1971). What I've said about Russell's view is influenced by some of Kripke's lectures on which his paper there is based.

'valence' from a large class of sentences containing it, and in a uniform way. For instance,

(5) For any elements $A$ and $B$, if one atom of $A$ combines with two of $B$, then the valence of $A$ is $-2$ times that of $B$.

is materially equivalent to

(5′) For any elements $A$ and $B$, if one atom of $A$ combines with two of $B$, then either $A$ is sodium and $B$ is sodium and $+1 = -2$ $(+1)$, or . . ., or $A$ is sulphur and $B$ is sodium and $-2 = -2$ $(+1)$, or . . .

provided that (3) is a correct and complete list of valences. So if anyone ever wants to eliminate the word 'denote' or the word 'valence' from a large class of English sentences by a uniform procedure, denotation definitions and valence definitions are just the thing he needs. There are, however, sentences from which these words are not eliminable by the sketched procedure. For instance, in semantics and possibly in chemistry there are problems with counterfactuals, e.g., 'If 'Germany' had been used to denote France, then . . .'. Moreover, there are special problems affecting the case of semantics, arising from the facts

(i) that the elimination procedure works only for languages in which nothing is denoted that cannot be denoted (without using semantic terms) in one's own language,
(ii) that it works only for languages that contain no ambiguous names,

and

(iii) that the denotation definitions provide no procedure for eliminating 'denote' from sentences where it is applied to more than one language; e.g., it gives no way of handling sentences like " 'Glub' denotes different things in different languages."

But, subject to these three qualifications (plus perhaps that involving counterfactuals), the elimination procedure for 'denote' is every bit as good as that for 'valence'!

What value did Tarski attach to such transformations? Unfortunately he did not discuss the one about valences, but he did discuss the one that transforms "Smith used a proper name to denote Germany" into something logically equivalent to "Smith uttered 'Deutschland'." And it is clear that to this definition he attached great philosophical importance. After defining semantics as "the totality of considerations concerning those concepts which, roughly speaking, express certain connexions between the expressions of a

language and the objects and states of affairs referred to by those expressions" (ESS 401), he says that with his definitions, "the problem of establishing semantics on a scientific basis is completely solved" (ESS 407). In other places his claims are almost as extravagant. For instance, the remark about physicalism that I quoted at the end of section II is intended to apply to denotation as well as to truth: if definitions of denotation like DE and DG could not be given, "it would . . . be impossible to bring [semantics] into harmony with . . . physicalism" (ESS 406); but because of these definitions, the compatibility of the semantic concept of denotation with physicalism is established. By similar standards of reduction, one might prove that witchcraft is compatible with physicalism, as long as witches cast only a finite number of spells: for then 'cast a spell' can be defined without use of any of the terms of witchcraft theory, merely by listing all the witch-and-victim pairs.

In other places Tarski makes quite different claims for the value of his denotation definitions. For example:

> We desire semantic terms (referring to the object language) to be introduced into the meta-language only by definition. For, if this postulate is satisfied, the definition of truth, or of any other semantic concept [including denotation, which Tarski had already specifically mentioned to be definable], will fulfill what we intuitively expect from every definition; that is, it will explain the meaning of the term being defined in terms whose meaning appears to be completely clear and unequivocal.[18]

But it is no more plausible that DE "explains the meaning of" 'denote' as applied to English, or that DG "explains the meaning of" 'denote' as applied to German, than that (3) "explains the meaning of" 'valence'—considerably *less* so in fact, since for 'valence' there is no analogue to the conclusions that 'denote' means something different when applied to English than it means when applied to German. In fact, it seems pretty clear that denotation definitions like DE and DG have no philosophical interest whatever. But what conclusions can we draw from this about Tarski's *truth* definitions like T2? I think the conclusion to draw is that *T2 has no philosophical interest whatever that is not shared by T1.* How this follows I will now explain.

We have seen that Tarski advocated theories of denotation for names that had the form of mere lists: examples of his denotation definitions were DE and DG, and for language L his denotation

[18] "The Semantic Conception of Truth and the Foundations of Semantics," *Philosophy and Phenomenological Research*, IV, 3 (March 1944): 341–375, p. 351.

definition would take the following form:

D2 $(\forall e)(\forall a)$ [$e$ is a name that denotes $a \equiv$ ($e$ is '$c_1$' and $a$ is $\bar{c}_1$) or ($e$ is '$c_2$' and $a$ is $\bar{c}_2$) or . . .]

where into the dots go analogous clauses for every name of L. Similarly, we can come up with definitions of application and fulfillment which are acceptable according to Tarski's standards, and which also have the form of mere lists. The definition of application runs:

A2 $(\forall e)(\forall a)$ [$e$ is a predicate that applies to $a \equiv$ ($e$ is '$p_1$' and $\bar{p}_1(a)$) or ($e$ is '$p_2$' and $\bar{p}_2(a)$) or . . .].

Similarly, we can formulate a list-like characterization F2 of fulfillment for the function symbols. Clearly neither A2 nor F2 is of any more theoretical interest than D2.

Tarski, I have stressed, accepted D2 as part of his semantic theory, and would also have accepted A2 and F2; and this fact is quite important, since D2, A2, and F2 together with T2 imply T1. In other words, T1 is simply a weaker version of Tarski's semantic theory; it is a logical consequence of Tarski's theory. Now, an interesting question is what you have to add to T1 to get the rest of Tarski's semantic theory. Suppose we can find a formula R that we can argue to be of no interest whatever, such that Tarski's semantic theory (T2 ∧ D2 ∧ A2 ∧ F2) is logically equivalent to T1 ∧ R. It will then follow that the whole interest of Tarski's semantic theory lies in T1—the rest of his semantic theory results simply by adding to it the formula R, which (I have assumed) has no interest whatever. And if there is nothing of interest in the conjunction T2 ∧ D2 ∧ A2 ∧ F2 beyond T1, certainly there can be nothing of interest in T2 alone beyond T1.

An example of such a formula R is D2 ∧ A2 ∧ F2: it is obvious that Tarski's semantic theory is logically equivalent to T1 ∧ D2 ∧ A2 ∧ F2. Because of this, *any interest in Tarski's semantic theory over* T1 *must be due to an interest in* D2 *or* A2 *or* F2 *(or to confusion): in this sense* D2 ∧ A2 ∧ F2 *is* "T2 *minus* T1". But I've already argued that D2, A2, and F2 have no theoretical interest whatever, and so that establishes that T2 has no theoretical interest whatever that is not shared by T1.

## V

Much of what I've said in this paper gains plausibility by being put in a wider persepctive, and so I now want to say a little bit about why we want a notion of truth. The notion of truth serves a great many purposes, but I suspect that its original purpose—the purpose for which it was first developed—was to aid us in utilizing the utter-

ances of others in drawing conclusions about the world. To take an extremely simple example, suppose that a friend reports that he's just come back from Alabama and that there was a foot of snow on the ground there. Were it not for his report we would have considered it extremely unlikely that there was a foot of snow on the ground in Alabama—but the friend knows snow when he sees it and is not prone to telling us lies for no apparent reason, and so after brief deliberation we conclude that probably there *was* a foot of snow in Alabama. What we did here was first to use our evidence about the person and his situation to decide that he probably said something true when he made a certain utterance, and then to draw a conclusion from the truth of his utterance to the existence of snow in Alabama. In order to make such inferences, we have to have a pretty good grasp of (i) the circumstances under which what another says is likely to be true, and (ii) how to get from a belief in the truth of what he says to a belief about the extralinguistic world.

If this idea is right, then two features of truth that are intimately bound up with the purposes to which the notion of truth are put are (I) the role that the attempt to tell the truth and the success in doing so play in social institutions, and (II) the fact that normally one is in a position to assert of a sentence that it is true in just those cases where one is in a position to assert the sentence or a paraphrase of it. It would then be natural to expect that what is involved in communicating the meaning of the word 'true' to a child or to a philosopher is getting across to him the sorts of facts listed under (I) and (II); for those are the facts that it is essential for him to have an awareness of if he is to put the notion of truth to its primary use (child) or if he is to get a clear grasp of what its primary use is (philosopher).

I think that this natural expectation is correct, and that it gives more insight than was given in sections II and IV into why it is that neither T1 nor T2 can reasonably be said to explain the meaning of the term 'true'—even when a theory of primitive reference is added to them. First consider (I). The need of understanding the sort of thing alluded to in (I), if we are to grasp the notion of truth, has been presented quite forcefully in Michael Dummett's article "Truth,"[19] in his analogy between speaking the truth and winning at a game. It is obvious that T1 and T2 don't explain anything like this (and in fact Dummett's fourth paragraph, on Frege-style truth definitions, can be carried over directly to T1 and T2).

[19] *Proceedings of the Aristotelian Society*, LIX (1958/9): 141–162.

The matter might perhaps be expressed in terms of assertibility conditions that one learns in learning to use the word 'true': part of what we learn, in learning to use this word, is that in cases like that involving the friend from Alabama there is some prima facie weight to be attached to the claim that the other person is saying something true. But there are also *other* assertibility conditions that one learns in learning the word 'true', assertibility conditions which have received considerable attention in the philosophical literature on truth. To begin with, let's note one obvious fact about how the word 'true' is standardly learned: we learn how to apply it to utterances of our own language first, and when we later learn to apply it to other languages it is by conceiving the utterances of another language more or less on the model of utterances of our own language. The obvious model of the first stage of this process is that we learn to assert all instances of the schema

(T) *X* is true if and only if *p*.

where '*X*' is replaced by a quotation-mark name of an English sentence *S* and '*p*' is replaced by *S*. This must be complicated to deal with ambiguous and truth-value-less sentences, but let's ignore them. Also let's ignore the fact that certain pathological instances of (T)—the Epimenides-type paradoxical sentences—are logically refutable. Then there is a sense in which the instances of (T) that we've learned to assert determine a unique extension for the predicate 'true' as applied to sentences of our own language.[20] Our views about what English sentences belong to this unique extension may be altered, but as long as we stick to the instances of (T) they cannot consistently be altered without also altering our beliefs in what those sentences express. This fact is extremly important to the functions that the word 'true' serves (as the Alabama example illustrates).

In stressing the assertibility conditions for simple sentences containing the word 'true', I have followed Quine (*ibid.* 138); for, like him, I believe that such assertibility conditions are enough to make the term 'true' reasonably clear. But now it might be asked, "Then why do we need causal (etc.) theories of reference? The words 'true' and 'denotes' are made perfectly clear by schemas like (T). To ask for more than these schemas—to ask for causal theories of reference to nail language to reality—is to fail to recognize that we are at sea on Neurath's boat: we have to work *within* our conceptual scheme, we can't glue it to reality from the outside."

[20] Cf. W. V. Quine, *From a Logical Point of View* (New York: Harper & Row, 1961), p. 136.

I suspect that this would be Quine's diagnosis—it is strongly suggested by §6 of *Word and Object*, especially when that is taken in conjunction with some of Quine's remarks about the inscrutibility of reference and truth value, the underdetermination of theories, and the relativity of ontology. It seems to me, however, that the diagnosis is quite wrong. In looking for a theory of truth and a theory of primitive reference we *are* trying to explain the connection between language and (extralinguistic) reality, but we are *not* trying to step outside of our theories of the world in order to do so. Our accounts of primitive reference and of truth are not to be thought of as something that could be given by philosophical reflection prior to scientific information—on the contrary, it seems likely that such things as psychological models of human beings and investigations of neurophysiology will be very relevant to discovering the mechanisms involved in reference. *The reason why accounts of truth and primitive reference are needed is not to tack our conceptual scheme onto reality from the outside; the reason, rather, is that without such accounts our conceptual scheme breaks down from the inside.* On our theory of the world it would be extremely surprising if there were some nonphysical connection between words and things. Thus if we could argue from our theory of the world that the notion of an utterer's saying something true, or referring to a particular thing, cannot be made sense of in physicalist terms (say, by arguing that any semantic notion that makes physicalistic sense *can* be explicated in Skinnerian terms, and that the notions of truth and reference *can't* be explicated in Skinnerian terms), then to the extent that such an argument is convincing we ought to be led to conclude that, if we are to remain physicalists, the notions of truth and reference must be abandoned. No amount of pointing out the clarity of these terms helps enable us to escape this conclusion: 'valence' and 'gene' were perfectly clear long before anyone succeeded in reducing them, but it was their reducibility and not their clarity before reduction that showed them to be compatible with physicalism.

The clarity of 'valence' and 'gene' before reduction—and even more, their *utility* before reduction—did provide physicalists with substantial reason to think that a reduction of these terms was possible, and, as I remarked earlier, a great deal of fruitful work in physical chemistry and chemical genetics was motivated by the fact. Similarly, insofar as semantic notions like 'true' are useful, we have every reason to suspect that they will be reducible to nonsemantic terms, and it is likely that progress in linguistic theory will come by looking for such reductions. (In fact, the fruitfulness of Tarski's work in aiding us to understand language is already some

sign of this, even though it represents only a partial reduction.) Of course, this sort of argument for the prospects of reducing semantic notions is only as powerful as our arguments for the utility of semantic terms; and it is clear that the question of the utility of the term 'true'—the purposes it serves, and the extent to which those purposes could be served by less pretentious notions such as warranted assertibility—needs much closer investigation.

All these remarks require one important qualification. The notion of valence, it must be admitted, is *not* reducible to nonchemical terms on the *strictest* standards of reduction, but is only *approximately* reducible; yet, in spite of this, we don't want to get rid of the notion, since it is still extremely useful in those contexts where its approximate character isn't too likely to get in the way and where if we did not approximate we'd get into quantum-mechanical problems far too complex for anyone to solve. (Moreover, considerations about the purposes of the notion of valence were sufficient to show that the notion of valence would only be approximately reducible: for the utility of the notion of valence is that it aids us in approximately characterizing which elements will combine with which and in what proportions; yet it is obvious that no *precise* such characterization is possible.)

Similarly, it may well be that a detailed investigation into the purposes of the notion of truth might show that these purposes require only an approximate reduction of the notion of truth. Still, to require an approximate reduction is to require quite a bit; after all, 'is a reincarnation of' isn't even approximately reducible to respectable biology, and 'electromagnetic field' is not approximately reducible to mechanics. Obviously the notion of approximate reduction needs to be made more precise (as in fact does the notion of strict, or nonapproximate, reduction); but even without making it so, I think we can see that T2 is no more of an approximate reduction than is V2, since D2 ∧ A2 ∧ F2 is no more of an approximate reduction than is (3). In other words, the main point of the paper survives when we replace the ideal of strict reduction by the ideal of approximate reduction.

It should be kept carefully in mind that the Quinean view that all we need do is clarify the term 'true', in the sense that this term is clarified by schema T (or by schema T plus a theory of translation to handle foreign languages; or by schema T plus the sort of thing alluded to in connection with Dummett), is *not* Tarski's view. Tarski's view is that we have to provide a truth characterization like T2 (which, when we choose as our object language L a "nice" fragment of our own language, can be shown correct merely by assuming

that all instances of schema T are valid—cf. fn 14, p. 361); and such a truth characterization does much more than schema T does. It does not do everything that Tarski ever claimed for it, for Tarski attached much too much importance to the pseudo-theories D2, A2, and F2; but even when we "subtract" such trivialities from his truth characterization T2, we still get the very interesting and important truth characterization T1. T1, I believe, adequately represents Tarski's real contribution to the theory of truth, and in doing this it has a number of positive advantages over T2 (in addition to the important negative advantage I've been stressing, of preventing extravagant claims based on the fact that T2 contains no semantic terms). First of all, T1, unlike T2, is applicable to languages that contain ambiguities and languages that contain terms not adequately translatable into English. Second, T1, unlike T2, can be used in diachronic linguistics: it doesn't need overhauling as you add new words to the language, provided those new words belong to the same semantic category as words already in the language. Third, I think that the reason why Tarski's theory of truth T2 has seemed so uninteresting to so many people is that it contains the vacuous semantic theories D2, A2, and F2 for the primitives of the language. By expressing the really important features of Tarski's results on truth, and leaving out the inessential and uninteresting "theories" of the semantics of the primitives, T1 should make the philosophical importance of Tarski's work more universally recognized.

HARTRY FIELD

Princeton University

## SOME OBSERVATIONS CONCERNING LOGICS AND CONCEPTS OF EXISTENCE *

I WILL begin by giving three reasons why we must think about the notion "concepts of existence" and about the relation between concepts of existence and logics. The rest of this paper puts together some ideas for thinking about these things. An intimate connection between concepts of existence, concepts of truth, and truth conditions (or "logical meanings") is uncovered. I try to show that certain aggregates of truth conditions generate logical syntaxes and that logical syntaxes so generated determine or project a concept of existence. It is then pointed out that Husserl's theory

* The author wishes to thank Michael Jubien, Paul Teller, Jeff Zucker, Street Fulton, and Robert Burch for their helpful remarks on earlier drafts.