Tarski, Truth, and Semantics

Richard G. Heck, Jr.

1 Opening

No one denies that Tarski made a major contribution to one particular problem about truth, namely, the resolution of the semantic paradoxes—although, of course, there is disagreement about whether he provided the *correct* solution. But some philosophers have suggested that Tarski also made a significant contribution to another project, that of providing semantic theories for natural languages. Hartry Field (2001), for example, credits Tarski with transforming the problem of reducing truth to physicalistically acceptable notions into that of reducing "primitive denotation". And Donald Davidson (1984c) founded an entire approach to semantics by arguing that a theory of meaning for a language may take the form of a Tarskian definition of truth.

But, according to John Etchemendy Etchemendy (1988),[1] in so far as Tarski's work does contribute to empirical semantics, this "is little more than a fortuitous accident". There are both conceptual and historical issues here. The conceptual question is whether reading Tarski's work on truth as it must be read, if it is to have any relevance to semantics, requires misunderstanding the character of his mathematical work. The historical question is whether Tarski *intended* his work to be so read. Etchemendy's view is that Tarski was primarily concerned to resolve the semantic paradoxes. Yet

> the form his solution takes appears... to serve equally as a characterization of the semantic properties of the language whose truth predicate is defined. However the appearance is actually quite misleading.... In particular, it would be a mistake to construe Tarski as taking part in this latter, semantic project, since the two goals turn out to be in quite direct opposition to one another. (Etchemendy, 1988, p. 52)

The form of Etchemendy's argument is thus this: the conceptual question must be answered affirmatively; the historical question should therefore be answered

[1] Others have argued for related claims. See Soames (1984) and Putnam (1994). I shall concentrate on Etchemendy's paper, since it contains the clearest, most complete, most influential presentation of the view.

negatively, on pain of charging Tarski with being just as confused as Davidson and Field are. For this reason, among others, I shall focus on the conceptual issue: for, if I am right that one does not have to be confused about what Tarski was doing to think he made a major contribution to formal semantics, Etchemendy's argument for the historical claim is also undermined.

The conceptual issue is also important in its own right. Davidson's original claim was that a theory of meaning may take the form of a Tarskian definition of truth. And although he has since modified his position, the role previously played by a definition now being played by a *theory* of truth, both Etchemendy and Davidson suggest that to abandon Tarski's definition of truth is to forfeit the guarantee of consistency it provides. But I shall argue that to think this is to be confused about the nature of Tarski's response to the liar paradox. More generally, I shall claim that there are at least two ways of reading Tarski's work so that it simultaneously contributes both to empirical semantics and to the resolution of the semantic paradoxes, in effect, by separating the mathematical from the empirical aspects of semantic theory. Once that has been established, the historical question will be easily answered, for it will then be clear that Tarski was fully aware that his work can be so read.

2 Tarski and the Problem of Consistency

By 1933, when Tarski published his famous paper on truth (Tarski, 1958a), the notion of truth had come to play a central role in logical theory. It is implicated in such notions as validity, logical consequence, completeness, and satisfiability. Informal mathematical theories involving these notions were widespread.[2] But Tarski saw, more clearly than anyone else, that these theories were threatened by inconsistency: the liar paradox, and the other semantic paradoxes, might prove to be derivable in them.[3] Tarski's goal, first and foremost, was to show that the notion of truth can be used in a consistent fashion—that it is possible to have a consistent, formalized mathematical theory in which the notion of truth has a place, and in which it can be used in the way it was already being used informally.

A central part of Tarski's response to this problem is his definition of truth,

[2] See, for example, Gödel (1986) Validity is there defined as satisfaction by all sequences: a sequence satisfies a formula if the result of substituting the members of the sequence for the relevant variables is true. The notion of satisfaction was thus not original to Tarski. His contributions were two: to do away with the obscure idea of substituting a set, or number, for a variable; and to use this notion to give a precise account of the semantics of the quantifiers.

[3] Tarski also saw that similar paradoxes threatened the notion of definability, and he worked on this problem, too (Tarski, 1958d,e). See also his remarks on Richard's paradox (Tarski, 1944, p. 346).

for a particular language, in an "essentially richer" meta-language. It is frequently said that Tarski offered an *explicit* definition of a truth-predicate, and there is an obvious sense in which this is true. But it can also be misleading. What Tarski actually offers, however, is a recursive definition. I shall not present the definition in rigorous detail: in particular, I shall simply ignore aspects which are not relevant to the questions at issue here, such as the formalization (or arithmetization) of syntax and the mathematical machinery required to speak of infinite sequences.[4]

The language of (first-order) arithmetic[5] contains the logical expressions '$\wedge$', '$\neg$', and '$\forall$'; an infinite collection of variables x_1, x_2, and so forth; non-logical primitives '0', 'S', '$+$', and '$\times$'; and parentheses. *A* recursive definition of denotation for this language may then be formulated as follows (Tarski, 1958a, p. 193):[6]

> A term v denotes the object z with respect to a sequence σ if and only if either:
>
> 1. v is '0' and z is 0,
> 2. v is x_i and z is $\sigma(i)$,
> 3. v is $\ulcorner St \urcorner$ and t denotes x w.r.t. σ and $z = Sx$,
> 4. v is $\ulcorner t + u \urcorner$ and t denotes x w.r.t. σ and u denotes y w.r.t. σ and $z = x + y$, or
> 5. v is $\ulcorner t \times u \urcorner$ and t denotes x w.r.t. σ and u denotes y w.r.t. σ and $z = x \times y$.

The recursive definition of satisfaction is then:

> σ satisfies C if and only if either:
>
> 1. C is $\ulcorner t = u \urcorner$ and the denotation of t w.r.t. σ = the denotation of u w.r.t. σ,
> 2. C is $\ulcorner A \wedge \mathrm{B} \urcorner$ and σ satisfies A and σ satisfies B,
> 3. C is $\ulcorner \neg A \urcorner$ and σ does not satisfy A, or

[4] Tarski himself remarks that the use of infinite sequences is inessential: finite sequences, which can be coded by natural numbers, will suffice (Tarski, 1958a, p. 195, fn. 1). This is important, for it is not the reference to *sequences* that mandates the use of an essentially richer meta-language.

[5] Tarski gives his definition for what he calls "the calculus of classes": I shall here silently transpose his discussion to the context of the language of arithmetic—in part, because it is more familiar; in part, because there are points I wish to make below which are most easily made in application to it.

[6] Notation: 'σ' and 'τ' range over infinite sequences of objects from the universe of discourse of the object-language (that is, natural numbers); '$\sigma(i)$' denotes the i^{th} member of σ.

4. C is $\ulcorner \forall x_i(A) \urcorner$ and every sequence τ differing from σ, if at all, only in the i^{th} place satisfies A.

And finally: A is true iff A is a sentence and every sequence satisfies A.

It is with this recursive definition that Tarski actually works. Nonetheless, as he mentions, given a mathematical theory of sufficient power, the recursive definition can be converted into an explicit definition by means of a general technique due to Frege and Dedekind (Tarski, 1958a, p. 193, fn. 1). The method works by converting the clauses of the recursive definition into conditions on sets: the intersection of all sets satisfying the conditions is then the extension of the predicate or function we seek to define. Thus, for example, we may define the notion of denotation as follows. Consider all sets D satisfying the following conditions:

$$< \text{‘0’}, 0, \sigma > \in D$$

$$< x_i, \sigma(i), \sigma > \in D$$

$$< t, x, \sigma > \in D \ \rightarrow < \ulcorner St \urcorner, Sx, \sigma > \in D$$

$$< t, x, \sigma > \in D \wedge < u, y, \sigma > \in D \ \rightarrow < \ulcorner t + u \urcorner, x + y, \sigma > \in D$$

$$< t, x, \sigma > \in D \wedge < u, y, \sigma > \in D \ \rightarrow < \ulcorner t \times u \urcorner, x \times y, \sigma > \in D$$

Note that these conditions are mere transcriptions of the clauses of the recursive definition (where ‘$< x, y, \sigma > \in D$’ is read as ‘x denotes y with respect to σ’). We then define Den as the intersection of all such sets. We may, in a similar way, transcribe the clauses in the recursive definition of satisfaction and produce an explicit definition of a set Sat. Finally, we define True to be the set of sentences A such that, for every (equivalently, for some) σ, $< A, \sigma > \in$ Sat.

It is easy to see that the predicate ‘$\in$ True’, so defined, meets Tarski’s condition of material adequacy: given any sentence of the language of arithmetic, one can prove a T-sentence for it, using these definitions.[7] The theory, as augmented by the definitions, is guaranteed to be consistent, if the theory in which the definition is formulated is consistent. (In fact, it will be a conservative extension of the original theory.) So the definition of truth shows that, if we distinguish object-language from meta-language, we can use the notion of truth in a consistent and materially correct fashion, and so the definition of truth resolves the problem that most exercised Tarski. Note, however, that the T-sentences come out as *theorems* of the meta-theory, in this case, say, of second-order arithmetic plus certain definitions: more precisely, the T-sentences are definitional transcriptions of theorems of second-order arithmetic; they arise from theorems of second-order arithmetic

[7] By a T-sentence I mean a sentence like: ‘$2+2=4$’ $\in$ True $\equiv 2+2=4$.

by substituting formulae of the form '$< x, y, \sigma > \in$ Den', '$< A, \sigma > \in$ Sat', and '$A \in$ True' for the arithmetical formulae that define them.

This is a point which it took some time to get across: that it is now so widely appreciated is due entirely to the heroic efforts of Etchemendy, Soames, Putnam, and others. But once it has been made, it follows immediately that Tarski's definition of truth is, by itself, no good as a semantic theory.[8] *No way is it a theorem of arithmetic* that '$2+2=4$' is true if and only if $2+2=4$; it's an empirical fact about what expressions in the language of arithmetic mean.[9]

How then can Tarski's work even *appear* as if it makes a real contribution to semantics? Consider a simpler example of a recursive definition, that of addition.[10] The recursive definition takes the following form:

$x+z$ is:

1. x, if $z=0$
2. $S(x+y)$, if $z=Sy$

There are two, quite different ways of embodying this recursive definition in a formal theory of arithmetic.[11] The first is to convert the recursive definition into an explicit one, by means of the techniques outlined above. Consider all sets A satisfying the following two conditions:

$$< x, 0, x > \in A$$

$$< x, y, z > \in A \rightarrow < x, Sy, Sz > \in A$$

Let Add be the intersection of all such sets; it is easy to see that $< x, y, z > \in$ Add if and only if $x+y=z$. This sort of definition is common in formal theories of second-order arithmetic.

[8] I am agreed with Etchemendy and Putnam, as against Soames, that it also follows that it is no good as a philosophical theory of truth. Whether Tarski intended to be offering any such theory is another matter. Burton Dreben has argued (in a lecture given at a colloquium at Boston University) that he did not, that his primary concern was to interpret mathematical talk of truth in set-theory. As I shall argue below, he most certainly did intend to do that: but his claiming to have reduced semantics to morphology suggests he had greater ambitions. Fortunately, I need take no stand on this question here. But, for what it is worth, even if Tarski did not provide us with a complete philosophical account of truth, his showing us how to formulate a consistent *theory* of truth is a major contribution.

[9] It is also frequently noted that the analysis gets the truth-values of certain counterfactuals wrong, for example: If 'snow is white' had meant that grass is pink, then 'snow is white' would have been false. This comes out false, because, on Tarski's definition, it is a *theorem* that 'snow is white' is true iff snow is white.

[10] This definition derives from Dedekind (1902, §135). Dedekind's definition differs from the one to be considered here in so far as he begins the series of natural numbers with one, not zero.

[11] Actually, there are three: some formal theories allow the introduction of new expressions by means of recursive definitions. For present purposes, however, such theories may be lumped with those which contain explicit definitions.

This approach is not the only one available, however, and it is not available at all unless the formal theory in which we are working is sufficiently strong. Alternatively, then, one may transform theclauses of the recursive definition into axioms governing a primitive, binary functional expression '+'.[12] Thus, in a formal theory of first-order arithmetic (for example, in PA), one typically finds the two axioms:

$$x + 0 = x$$
$$x + Sy = S(x + y)$$

These are the so-called recursion equations for addition.

Now Tarski, as we have seen, offered recursive definitions of denotation and satisfaction. And just as in the case of addition, there are two ways the definition can be embodied in a formal theory. If the theory is sufficiently strong, the recursive definition can be converted into an explicit one, as we saw above. Alternatively, one can transform the clauses of the recursive definition into axioms in a formaltheory of truth. Thus, for example, such a theory might have the following ten axioms:[13]

1. $\mathrm{den}(\text{`0'}, \sigma) = 0$
2. $\mathrm{den}(x_i, \sigma) = \sigma(i)$
3. $\mathrm{den}(\ulcorner St \urcorner, \sigma) = S(\mathrm{den}(t, \sigma))$
4. $\mathrm{den}(\ulcorner t + u \urcorner, \sigma) = \mathrm{den}(t, \sigma) + \mathrm{den}(u, \sigma)$
5. $\mathrm{den}(\ulcorner t \times u \urcorner, \sigma) = \mathrm{den}(t, \sigma) \times \mathrm{den}(u, \sigma)$
6. $\mathrm{sat}(\ulcorner t = u \urcorner, \sigma) \equiv \mathrm{den}(t, \sigma) = \mathrm{den}(u, \sigma)$
7. $\mathrm{sat}(\ulcorner A \wedge \mathrm{B} \urcorner, \sigma) \equiv \mathrm{sat}(A, \sigma) \wedge \mathrm{sat}(B, \sigma)$
8. $\mathrm{sat}(\ulcorner \neg \mathrm{A} \urcorner, \sigma) \equiv \neg \mathrm{sat}(A, \sigma)$
9. $\mathrm{sat}(\ulcorner \forall x_i(A) \urcorner, \sigma) \equiv \forall \tau[\forall j(i \neq j \rightarrow \sigma(j) = \tau(j)) \rightarrow \mathrm{sat}(A, \tau)]$
10. $\mathrm{true}(A) \equiv$A is a sentence $\wedge \forall \sigma(\mathrm{sat}(A, \sigma))$

[12] This important idea first appears in Skolem (1967). By the time of Tarski's work on truth, it had become well-known.

[13] Notation: read '$\mathrm{den}(t, \sigma)$' as 'the denotation of t w.r.t. σ'; '$\mathrm{sat}(A, \sigma)$', as 'A is satisfied by σ'; '$\mathrm{true}(A)$', as 'A is true'.

I have heard it said that a theory of this form can not be finitely axiomatized, due to the need to have a special axiom for each of the variables. But this is a mistake. The second axiom should be read as follows: for every n, the n^{th} variable denotes, with respect to a sequence σ, the n^{th} member of σ. Similar remarks apply to the axiom governing the universal quantifier.

It is easy to see that this theory is also materially adequate in Tarski's sense.

An axiomatic theory of truth, unlike the explicit definition of truth considered above, treats the notions of truth, denotation, and satisfaction as undefined primitives: for that reason, it may properly be thought of as a formalization of an *empirical* semantic theory. Of course, the T-sentences are theorems of this theory, too, but that is to say no more than that they are derivable from the axioms—and if the axioms have empirical content, there is no reason the theorems shouldn't also have empirical content. It is thus no more mysterious how Tarski's work came to have an influence on formal semantics than how Dedekind's definition of addition gave rise to the usual axioms for addition.

But the question remains whether one does not, in effect, misunderstand what *Tarski* was trying to do if one formalizes his recursive definition of truth as an axiomatic theory. Etchemendy, I expect, would want to claim that so construing it amounts to a gross misunderstanding of Tarski's project: in particular, merely presenting a formal, axiomatic theory of truth does not show that it is possible to use the notion of truth consistently. The theory does respect the distinction between object-language and meta-language—there is no axiom that tells us anything about expressions of the form 'true(A)'—and that might give us reason to *hope* that the theory is consistent. Still, the axiomatic theory on its own does not solve Tarski's problem.

3 The Connecting Principles

In order to answer the questions with which we began, it should suffice to understand the relationship between the axiomatic theory—which embodies Tarski's disputed contribution to empirical semantics—and the explicit definition of truth—which is essential to his claim to have proven that the notion of truth can be consistently employed in meta-mathematics. Etchemendy sees the situation in much the same way, and he offers an interesting account of this relationship: "... if we define a set TRUE using the standard recursive definition, then the claim that all and only the true sentences of the language are members of TRUE is logically equivalent to the" axiomatic theory outlined above (1988, p. 59).[14] That is to say, suppose we add primitive expressions 'true', 'den', and 'sat' to the language

[14] Etchemendy's claim that the theories are "logically equivalent" is somewhat imprecisely stated. Whether the theories are equivalent will depend upon the strength of the theory in which the definition is given: if we give a definition of truth for the language of arithmetic in ZFC, the axiomatic theory outlined above is certainly not equivalent to ZFC. However, in most cases, a theory will have a natural extension in which truth can be defined and which will be equivalent to the axiomatic theory of truth for the original theory. The mathematical situation is subtle, and these sorts of results can be very sensitive to how the respective theories are formulated. See Feferman (1977).

in which the definition is given and adopt three new axioms, which are intended to express the claims that the set True contains all and only the true sentences of the object-language and that Den and Sat are, respectively, the extensions of the denotation-function and satisfaction-relation:

$$\text{true}(A) \text{ iff } A \in \text{True}$$

$$\text{den}(t,\sigma) = x \text{ iff } < t,x,\sigma > \in \text{Den}$$

$$\text{sat}(A,\sigma) \text{ iff } < A,\sigma > \in \text{Sat}$$

Then all the axioms of the axiomatic theory can be proven from the explicit definition of truth and these new axioms, the "connecting principles".

As Etchemendy puts the point: "... thanks to the techniques Tarski uses in his definition, the claim that all and only the true sentences are members of the defined set takes on genuine semantic import" (1988, p. 60). One should, therefore, wonder how he can claim that it is but an accident that Tarski's work on truth has any relevance to semantics. The answer is that Etchemendy wants to claim, not only that Tarski would not have endorsed the connecting principles, but that they "are most emphatically not part of Tarski's project, but in an obvious sense conflict with it, involving as they do the uneliminated use of a notion of truth" (1988, p. 60).

Now, there is an obvious sense in which Tarski made no such claims as those embodied in the connecting principles: the formal machinery of the definition does not make use of primitive notions of truth, denotation, and satisfaction. But there is more to Tarski's work on truth than his definition: there is, in particular, his claim that the definition is materially adequate. It is easy to overlook the significance of this point, since the condition of material adequacy Tarski states in "The Concept of Truth" is a formal one, that the definition should enable us to prove all T-sentences for sentences of the object-language. But it is clear that the *point* of the condition is to guarantee that the predicate he defines has the right extension—that it is a *truth*-predicate rather than a falsity-predicate or a well-formedness predicate. To say that the condition is adequate "materially", as well as formally, is to say that it not only avoids inconsistency (and is otherwise acceptable, mathematically speaking), but that it gets the facts right.[15] Tarski can state the condition of material adequacy as the formal condition he does because, first, any two predicates

[15] See (1958e, pp. 128–9), where he takes material adequacy to be an even stronger condition: "Now the question arises whether the definitions just constructed... are also adequate materially; that is, do they in fact grasp the current meaning of the notion as it is known intuitively? Properly understood, this question contains no problem of a purely mathematical nature, but it is nevertheless of capital importance for our considerations."

meeting the formal condition must have the same extension;[16] and secondly, any predicate that expresses the intuitive notion of truth—say, the English predicate 'true'—will itself meet this condition. It follows that, if '$\in$ True' meets the formal condition of material adequacy, it must have the same extension as 'true' and so will have the *right* extension.

Now, what the connecting principle concerning truth says is precisely that Tarski's truth-predicate has the same extension as the predicate 'true', and similar remarks apply to the other connecting principles. So the connecting principles express that the definitions of truth, denotation, and satisfaction are materially adequate, claims one can hardly describe as "not part of Tarski's project". As said, the whole point of the notion of material adequacy is to guarantee the truth of the connecting principles.

There is, of course, a difference between the condition of material adequacy, on the one hand, and the connecting principles, on the other: the former can be formulated without using the intuitive notion of truth, whereas the latter can not. And so, Etchemendy might say, the connecting principles can not be of any relevance to Tarski's project, since their statement requires the use of a primitive notion of truth. One might well respond that the condition of material adequacy can not be formulated, for the general case, without appeal to the notion of translation (which is at least as problematic as that of truth). But however that may be, what began as a correct point about Tarski's resolution of the semantic paradoxes has now become something quite different. True enough: Tarski wanted to provide an eliminative definition of the truth-predicate because it was by doing so that he intended to show that the notion of truth can be used in a consistent and materially correct fashion. But it is hard to see why this should imply that he had a *general* objection to the introduction of a primitive notion of truth; what he did object to was any introduction of such a notion that brings with it a threat of inconsistency.[17] And it is obvious that

[16] On sentences of their common object-language, of course. For if, in general, S is true iff p, and S is schmue iff p, then S is true iff S is schmue. See Tarski (1958a, p. 258) where he requires, as a condition on the adequacy of a definition of truth, that the connecting principles should be provable if another definition of truth is added.

[17] The only relevant remarks known to me are at Tarski (1958b, pp. 405–6). Tarski makes a series of objections to the use of axiomatic theories of truth. One is that the consistency of the theory needs proving—but this means simply that more mathematical work is needed (and so, in particular, that we still need the definition of truth, which seems to be his main point). Another is that "the choice of axioms always has a rather accidental character, depending on inessential factors (such as e.g. the actual state of ourknowledge)". The objection may seem obscure: but the same remark is made, almost *verbatim*, during Tarski's discussion of languages of infinite order (1958a, p. 258). The claim is not that the axioms of a *structured* axiomatic theory are arbitrary, but that such a situation afflicts the *minimal* theory of truth for such a language: we should have to add generalizations about truth one by one; which we decided to add would depend upon the state of our knowledge of *semantics* (see the discussion of such theories below). And, infamously, Tarski remarks that an axiomatic

the connecting principles pose no such threat: formally, they can be understood as definitions, as mere re-writings of '$A \in$ True' as 'true(A)', and so forth. They need not be so interpreted, but they pose no more danger than definitions would.

Tarski repeatedly claims to have reduced semantics to something like morphology (Tarski, 1958a, pp. 251–4; 1958b, p. 406). I think he over-states his accomplishments. What he has done is cleanly to separate the mathematical from the empirical aspects of semantics: the mathematical part is wholly absorbed into the definition of truth; the empirical, into the claim of material adequacy. Formally, this separation is crucial: it makes a purely mathematical treatment of truth possible, at least for certain purposes. And although the connecting principles *could* be added to the theory in which the definition of truth is given, with no threat of inconsistency, there would be little point in this: they would assert that the definition is materially adequate, but they can hardly be proven, so the claim of material adequacy might just as well be left at an informal level.

4 The Relationship Between the Axiomatic Theory and Tarski's Definition of Truth

One might therefore think of Tarski as having contributed to semantics as follows: he showed that one can formulate a theory of truth with genuine semantic interest by giving an explicit definition of truth for the object-language and appending the connecting principles.[18] The reason this would be a contribution to semantics is, as Etchemendy says, that the axioms of the axiomatic theory would then be derivable: the semantic interest of the definition plus the connecting principles is thus parasitic on that of the axiomatic theory. This fact makes the question whether Tarski would have endorsed the connecting principles moot: for, even if he would not, one could yet understand Tarski has having contributed to semantics by offering a recursive definition of truth, which can be formalized not only as an explicit definition in a stronger meta-theory, but also as an axiomatic theory of truth.

In any event, it is as an axiomatic theory of truth that an empirical semantic theory would most likely be formulated. Etchemendy, however, sees any step away from a *definition* of truth as momentous:

> We have seen that to do semantics we must reintroduce a primitive notion of truth: a Tarskian truth predicate [one that is recursively or

theory might prove "difficult to bring... into harmony with the postulates of the unity of science and of physicalism...". But I do not intend to add to existing discussion of this claim.

[18] Remarkably enough, something like this seems to have been Davidson's view at one point. For he once held that an interpreter should have a definition of truth for his target and know *of* that definition that it satisfies various empirical constraints, that it is "interpretive", which is Davidson's replacement for the notion of material adequacy (Davidson, 1984b, p. 139).

> explicitly defined] does not allow us to make the substantive semantic claims. . . that constitute, in part, the goal of semantics. Now there is a certain irony here that should not go unmentioned. For although Tarski provides a solution to the semantic paradoxes usable in a wide range of situations, *that solution is specifically not available to those doing semantics*. (Etchemendy, 1988, p. 64, my emphasis)

The worry would seem to be this: Tarski's solution is essentially connected with his explicit definition of truth; but an explicitly defined truth-predicate is no good for semantics. In order to do semantics, we must make use of a primitive notion of truth, whence the assurance of consistency provided by Tarski's explicit definition lapses.

This argument is extremely problematic. We have already seen that there is at least one way to "reintroduce a primitive notion of truth" which brings with it no threat of inconsistency: tack the connecting principles onto a definition of truth. But, as just said, this approach to semantics may not be best: we may want to abandon Tarski's definition of truth and formulate our semantic theories axiomatically. This is probably what Etchemendy means by "doing semantics". And the claim is that we must renounce Tarski's solution to the semantic paradoxes, if we want to do semantics in this sense.

This last claim rests upon a misunderstanding of what Tarski's solution is. In particular, it presupposes that the solution consists in renouncing the use of any but an explicitly defined notion of truth. But this does not fit Tarski's diagnosis of the paradoxes. According to Tarski, the liar paradox will arise in any theory which is "semantically closed", a theory being semantically closed if it proves T-sentences for all sentences of its language.[19] Tarski proves that any such theory is inconsistent, and his response to the liar paradox is "*not to use any language which is semantically closed*" (Tarski, 1944, p. 349; Tarski's italics):[20] that is, the

[19] Tarski also requires that the langauge contain quotation-names of all its sentences (something which all sufficiently strong languages will do, *via* arithmetization) and that "the ordinary laws of logic" (that is, of classical logic) hold (Tarski, 1944, p. 348; 1958a, p. 165). Saul Kripke (1975) abandons this last assumption. Of course, he also dethrones the T-sentences, but, in a theory with a three-valued logic of the usual sort, this move is independently motivated: a sentence can not satisfy its T-sentence unless it is either true or false. The T-*rules*—rules of inference which allow us to pass between A and $\ulcorner\mathrm{True}(\ulcorner\mathrm{A}\urcorner)\urcorner$, and between $\ulcorner\neg\mathrm{A}\urcorner$ and $\ulcorner\neg\mathrm{True}(\ulcorner\mathrm{A}\urcorner)\urcorner$—are the natural analogue of the T-sentences here. In Kripke's theory, the T-rules hold for all sentences of the language, *including* those that contain 'True'. What is distinctive about Kripke's theory of truth is that no distinction between object-language and meta-language need be made *within* it.

[20] Tarski is quite sloppy about the distinction between languages and theories formulated in those languages. Languages are not inconsistent, and they do not imply things, so it is not the *language* that is semantically closed, but some theory formulated in that language. The point is important in evaluating Tarski's suggestion that natural languages, such as English, are inconsistent. The remark,

solution is to enforce the distinction between object-language and meta-language, and so to require only that the theory of truth should imply T-sentences for all sentences of the *object*-language.[21]

One can, of course, respect the distinction between object- and meta-language in an axiomatic theory of truth. But that means that Tarski's solution *is* available to those "doing semantics". As said above, that might give us reason to hope that the axiomatic theory discussed above is consistent: but the worry is that, if we abandon the explicit definition of truth, then, even if Tarski's solution is still available, his work provides no *assurance* of consistency. Davidson, for example, is clearly troubled by this thought. Recognizing that a Tarskian definition of truth is no good for semantics, he proposes, much as I have above, that we formalize Tarski's recursive definition of truth as an axiomatic theory. But, he says, it would appear that "only an explicit definition can guarantee the consistency of the resulting system". Davidson remarks, seemingly not very satisfied himself, that this objection can be "evaded as long as known ways of producing paradox are not introduced" (Davidson, 1990, p. 297, fn. 34). But this is too weak: the objection can be refuted, and it is Tarski who shows us how.

One way of proving the consistency of one theory (the target theory) relative to another (the base theory) is as follows: define the primitives of the target theory in terms of those of the base theory; then, prove the axioms of the target theory in the base theory, augmented by the definitions. This having been done, the target theory is said to have been interpreted in the base theory.[22] It then follows that the target theory is consistent if the base theory is: if one could derive a contradiction from the axioms of the target theory, one could mimic that derivation in the base theory plus the definitions by proving the necessary axioms of the target theory and appending the derivation of a contradiction, in the target theory, from them.

as it stands, is not even coherent. See Boolos (1975).

[21] It is sometimes suggested that sentences like "'2+2 = 4" is true" is true' are not even well-formed for Tarski. This is wrong, at least as regards formal theories of truth. The syntax of the meta-language can be arithmetized just like the syntax of the object-language, and one can then formulate such sentences in the meta-language. In the intended model of the language, some such T-sentences will be true (and the meta-theory can even be strengthened so that some such T-sentences are provable). But the crucial point is that not all such T-sentences will be true: In particular, the T-sentence for the liar sentence will not be true (and any strengthening of the meta-theory that makes it provable will be inconsistent).

[22] The notion of interpretation was not made mathematically precise until some twenty years after the publication of "The Concept of Truth"—interestingly enough, by Tarski himself (1953). A relative interpretation of one theory in another meets similar conditions, but one is also allowed to relativize the quantifiers of the base theory by means of some formula, i.e., restrict the domain of the quantifiers appearing in the target theory. In some cases, too, one needs to show that rules of inference of the target theory are derived rules of the base theory. But these complications need not detain us.

Tarski's work on truth can be read in this light. Here, the target theory is an axiomatic theory of truth for, say, the language of arithmetic; the base theory is second-order arithmetic.[23] The connecting principles may be construed as definitions of the primitives of the axiomatic theory in second-order arithmetic. Now, as Etchemendy remarks, all axioms of the axiomatic theory are provable from the definition of truth plus the connecting principles. That establishes that the axiomatic theory can be interpreted in second-order arithmetic, and so that the axiomatic theory of truth is consistent if second-order arithmetic is. Since the connecting principles are little more than re-statements of the conditions of material adequacy, it would be at best uncharitable not to credit Tarski with having proved as much.

We therefore need not worry that, if we abandon Tarski's explicit definition of truth in favor of an axiomatic theory of truth, we can only profess faith in the distinction between object- and meta-language when asked why we do not fear inconsistency. On the contrary, Tarski shows us how to assure ourselves of the consistency of *any* such axiomatic theory of truth, so long as it respects the distinction between object-language and meta-language. Here is the recipe: reconstrue the axioms of the theory as clauses in recursive definitions of denotation and satisfaction; use standard techniques to convert these recursive defninitions into explicit ones in an appropriate meta-theory; note that the axioms of the axiomatic theory are consequences of the explicit definitions and the connecting principles, themselves reconstrued as definitions of the primitives of the axiomatic theory in the meta-theory; finally, note that all of this amounts to an interpretation of the axiomatic theory in the meta-theory. Of course, if the meta-theory is inconsistent, this won't be of much help, and the best we can hope to do is to find a meta-theory which is proof-theoretically equivalent to the axiomatic theory. But Gödel taught us that long ago, and we shall just have to live with it.

Thus, to summarize: Tarski's solution to the semantic paradoxes is not to provide an explicit definition of truth; his solution is to distinguish object-language from meta-language. The explcit definition is part of Tarski's proof that his solution *works*: that, if we respect this distinction between object-language and meta-language, the axiomatic theory of truth he shows us how to construct will be consistent, so long as the theory in which the explicit definition of truth is given is consistent.

[23] The base theory may, in fact, be much weaker. Certainly Π^1_1 second-order arithmetic suffices, and even this can be weakened.

5 Is the Recursive Character of Tarski's Definition Essential?

If Tarski made a real contribution to semantics, he did so by offering a recursive definition of truth for a particular language, which definition can be formalized as an axiomatic theory, and then proving that this theory is consistent (if some appropriate meta-theory is consistent). But, according to Etchemendy, Tarski could have solved the problem which concerned him—could have proven that the notion of truth can be used in a consistent and materially correct fashion—by giving a "list-like" definition of truth (Etchemendy, 1988, p. 60). For a language with only finitely many sentences, for example, one could define 'S is true' as follows:

> S is true iff $[(S = S_1)$ and $p_1]$ or $[(S = S_2)$ and $p_2]$ or $\ldots$ or $[(S = S_n)$ and $p_n]$

Such a definition can obviously be made to yield a truth-predicate that is materially adequate: just make sure the p_i are properly chosen. Moreover, the fact that the predicate had been explicitly defined would resolve the problem of consistency. But does it not then follow that the recursive character of Tarski's definition is inessential? that this machinery enters only because the language of arithmetic contains infinitely many sentences, and so no list-like definition is available? If so, the semantic interest of Tarski's work on truth is due to inessential features of the definition he gives, features the definition *need not have had*, given his goals. And one might well dramatize the situation by saying that his having contributed to semantics was but "a fortuitous accident".[24]

Whether Tarski would have been content with a list-like definition depends upon what purpose he wanted his definition of truth to serve.[25] As said earlier, Tarski's goal was not *just* to develop a consistent mathematical theory in which the notion of truth has a place, but a theory strong enough to prove theorems whose statements involve the notion of truth, or whose informal proofs had made use of it. What sorts of results are in question here is is clear from what Tarski goes on to do after he gives his definition of truth. Thus, he proves various 'generalizations about truth', such claims as the law of bivalence:

> For all sentences A, either $A \in \text{True}$ or $\ulcorner\neg A\urcorner \in \text{True}$

Moreover, he proves that any (deductive) consequence of a set of true sentences is true; that the set of consequences of any set of true sentences contains only

[24] That Tarski thought there were *some* advantages to the recursive definition is clear enough: he remarks that it "bring[s] out the content of the concept defined more clearly than the [explicit] definition does" (1958a, p. 177, fn. 1). But Etchemendy could concede that point.

[25] The remarks in this and the next two paragraphs were inspired by conversations with George Boolos, to whom I owe a large debt at this point.

true sentences; that the set of true sentences is complete and consistent; that every axiom (and so theorem) of the calculus of classes is true; and so that this theory is consistent, although incomplete (Tarski, 1958a, pp. 197–9). These results rest upon theorems of a more fundamental character, such as:

$$\text{For any sentences } A \text{ and } B, \ulcorner A \wedge B \urcorner \in \text{True} \quad \equiv \quad A \in \text{True} \wedge B \in \text{True}$$

That is to say, the results Tarski proves *depend upon the derivability of the axioms of the axiomatic theory from the definition of truth*. But these axioms do *not* follow from just any theory of truth that yields all the T-sentences.[26]

The point is worth a digression, for it bears directly upon so-called deflationary theories of truth. Paul Horwich has argued that what he calls the "minimal" theory of truth suffices to account for all uses of the truth-predicate. The minimal theory for a particular language contains as axioms all T-sentences for sentences of that language. Horwich argues that, on the basis of this theory, one can prove such facts as that, for any sentences A and B, if $\ulcorner A \rightarrow B \urcorner$ is true and A is true, then B is true (1990, p. 23).[27] But it can be proven that this does not follow from the minimal theory for, say, the language of arithmetic: there are models in which the T-sentences all hold, but in which this claim does not. In fact, none of the axioms of the axiomatic theory follow from the minimal theory for the language of arithmetic.[28]

[26] Etchemendy recognizes this fact (1988, p. 60) but apparently overlooks its significance. He might respond that the recursive clauses *will* be provable from the list-like definition if the language has only finitely many sentences—or that, in that case, we do not need to prove the recursive clauses to get results like those Tarski mentions. Still, we do need the recursive clauses, if the language has infinitely many sentences, which any interesting language will, and they will not be provable from the minimal theory in that case. Since Tarski was interested in languages of this sort, for *his* purposes, a list-like definition will not do.

[27] In fact, Horwich attempts to prove something different: that if A implies B and A is true, then B is true. But if we can not prove what is mentioned in the text, we can hardly hope to prove this claim. Horwich's discussion is also formulated in terms of a notion of truth applying to propositions rather than to sentences. But this difference does not matter for present purposes.

[28] The problem is that the instances of the T-schema fix the extension of the truth-predicate only for the *standard* numbers, that is, for the Gödel numbers of standard sentences. We will be free to fix its extension on non-standard numbers as we wish, so there can be non-standard sentences that fail to satisfy such general laws of truth as that mentioned in the text.

Horwich's argument goes wrong at a number of points. The most obvious is that it makes use of free variables ranging over propositions, and most of the expressions occurring in his "proof" are not even well-formed. For example, one finds such expressions as '$< p >$ is true iff p'. But this makes no more sense if 'p' ranges over propositions than it would if it ranged over sentences. The most charitable reading would have Horwich offering not an argument but an argument-schema. His discussion does establish that every *instance* of

If $\ulcorner A \rightarrow B \urcorner$ is true, then, if A is true, then B is true

Tarski himself makes a similar observation in connection with languages of infinite order. His Theorem III states that the minimal theory of truth for such a language is consistent. But, he notes, the resulting theory "would be a very incomplete system, which would lack the most important and most useful theorems", for example, the generalizations about truth (Tarski, 1958a, p. 257). So more must go into the definition of truth, *if it is to do the work Tarski wants it to do*, than enough to enable us to prove all the T-sentences—and more, therefore, than an infinite list of them. The recursive structure of the definition, in particular, would appear to be necessary[29] if the truth-predicate Tarski defines is to be sufficient for meta-mathematics. So Tarski's purpose would not have been served by a list-like definition, the recursive character of the definition is essential, and his work's importance for semantics is no accident.

6 Closing: What Did Tarski Know? And When Did He Know It?

Tarski made *two* major contributions to semantics: he showed us how to formulate axiomatic theories of truth for various languages, and he proved that such theories are consistent, so long as they respect the distinction between object-language and meta-language. The conceptual question—whether reading Tarski's work as making major contributions to semantics requires misunderstanding the mathematics itself—is thus to be answered negatively. What of the historical question, then, whether Tarski intended to make a contribution to semantics? The interesting question is not whether he intended to contribute to semantics. There is little, if any, controversy about that: everyone agrees that Tarski wanted to establish a "scientific semantics".[30] The interesting question is whether Tarski knew that his work

follows from the minimal theory. But this is quite a different matter from showing that the generalization itself follows from the minimal theory.

This confusion—between the minimal theory's ability to prove every instance of a generalization and its ability to prove the generalization itself—is of general significance. As was pointed out to me by Tom Kelly, a similar situation obtains regarding the minimal theory's ability to explain the sorts of psychological generalizations central to the so-called "success argument" against deflationism. See here Gupta (1993, p. 67).

[29] Although not sufficient: there are theories in which the axioms (and so the T-sentences) can all be proven, but in which such general facts about truth as that no sentence is both true and false can not be proven. The simplest example is an axiomatic theory of truth for PA, in which the predicate 'true' and the other semantical predicates are not allowed to occur in the induction axioms.

[30] Of course, Tarski may not have meant by 'semantics' what we call semantics, and his thinking one so much as could reduce semantics to morphology might be taken to show this. On the other hand, he (and others) may just have been confused about a simple, but subtle, point. Tarski showed that we can can 'do semantics in set theory' in exactly the sense that Gödel showed that we can 'do syntax in arithmetic': the former can be interpreted in the latter. That no more shows that semantics can be *reduced* to set theory than showing that the theory of general relativity can be interpreted in

contributed that which, as it happens, has been of such significance for semantics. That he did follows from the fact that *he* reads his work just as I have suggested *we* should read his work.

In discussing languages of infinite order, Tarski writes that, since it is impossible to give an explicit definition of truth for such languages,[31] we must try another way:

> The idea naturally suggests itself of setting up semantics as a special deductive science with a system of morphology as its logical substructure. For this purpose it would be necessary to introduce into morphology a given semantical notion as an undefined concept and to establish its fundamental properties by means of axioms. The experience gained from the study of semantical concepts in connexion with colloquial language, warns us of the great dangers that may accompany the use of this method. For that reason the question of how we can be certain that the axiomatic method will not in this case lead to inconsistencies becomes especially important. (Tarski, 1958a, p. 255)

The axiomatic theory he considers is, as was said earlier, the minimal theory of truth for a language of infinite order, and the problem is to prove that this theory is consistent. Suppose, then, that there were a proof of a contradiction in this theory. That proof could use only finitely many of the axioms of the minimal theory: hence, all of the axioms used would belong to the minimal theory of truth for a language of some finite order, say, k. So it will suffice to show that any such theory is consistent.[32] Tarski's argument for that claim is as follows:

> ... [A] definition of the symbol 'Tr' can be constructed in the metatheory such that the axioms [of the minimal theory for a language of order k] become consequences of this definition. In other words: these axioms, with a suitable interpretation of the symbol 'Tr', become provable sentences of the metatheory.... (Tarski, 1958a, pp. 256–7)

That is, the consistency of the minimal theory of truth for a language of finite order k follows from the fact that the explicit definition of truth constitutes an interpretation of that theory in a meta-theory of order k+1 (itself assumed to be

set-theory shows that *it* can be reduced to set theory. The empirical content of semantic theories does not vanish when they are formulated as explicit definitions of a truth-predicate: it is shifted, *in toto*, onto the claim that the theory is materially adequate.

[31] As Tarski soon realized, this is mistaken: see the Postscript to Tarski (1958a). But this does not matter at present.

[32] Tarski's argument is slightly different from the one presented here, but these differences do not matter for present purposes.

consistent). Similar considerations obviously show that the structured, axiomatic theories of truth we considered above are also consistent.[33]

Tarski was thus perfectly well aware that he had done what semanticists have long thanked him for doing. Of course, he may not have known that these particular aspects of his work would prove so important to semantics—and, given that he thought he had reduced semantics to morphology, he might have been appalled. But that does not relieve us of our debt to him.[34]

References

Boolos, G. (1975). The appearance of truth. Unfinished manuscript.

Davidson, D. (1984a). *Inquiries Into Truth and Interpretation*. Oxford, Clarendon Press.

—— (1984b). 'Radical interpretation', in Davidson 1984a, 125–139.

—— (1984c). 'Truth and meaning', in Davidson 1984a, 17–36.

—— (1990). 'The structure and content of truth', *Journal of Philosophy* 87: 279–328.

Dedekind, R. (1902). 'The nature and meaning of numbers', tr. by W. W. Beman, in *Essays on the theory of numbers*. Chicago, The Open Court Publishing Company, 31–115.

Etchemendy, J. (1988). 'Tarski on truth and logical consequence', *Journal of Symbolic Logic* 53: 51–79.

Feferman, S. (1977). 'Theories of finite type related to mathematical practice', in J. Barwise (ed.), *A Handbook of Mathematical Logic*. New York, North-Holland Publishing, 913–72.

Field, H. (2001). 'Tarski's theory of truth', in *Truth and the Absence of Fact*. Oxford, Clarendon Press, 3–26. reprinted, with a postscript (pp. 27-29).

[33] Tarski does not mention this, but it is hard to believe he would not have realized it. In the context of his discussion of languages of infinite order, the point would not arise, because the problem Tarski did not know how to solve—and, at that time, believed unsolvable—was how to *formulate* a structured axiomatic theory of truth for languages of infinite order.

[34] Thanks is due to George Boolos, Michael Glanzberg, Warren Goldfarb, Tom Kelly, Charles Parsons, Steve Peterson, and Jason Stanley for helpful conversations, as well as to an anonymous referee for the *Philosophical Review*. Thanks also to the students in my courses on truth and on the philosophy of language, given at Harvard University in the Springs of 1995 and 1996, respectively, in which I presented some of this material.

Gödel, K. (1986). 'The completeness of the axioms of the functional calculus of logic', in S. Feferman, *et al*. (eds.), *Collected Works*, volume 1, 3d edition. Oxford, Oxford University Press, 102–23.

Gupta, A. (1993). 'A critique of deflationism', *Philosophical Topics* 21: 57–81.

Horwich, P. (1990). *Truth.* Oxford, Blackwell.

Kripke, S. (1975). 'Outline of a theory of truth', *Journal of Philosophy* 72: 690–716.

Putnam, H. (1994). 'A comparison of something with something else', in *Words and Life*. Cambridge MA, Harvard University Press, 330–50.

Skolem, T. (1967). 'The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains', in J. van Heijenoort (ed.), *From Frege to Gödel: A Sourcebook in Mathematical Logic*. Cambridge MA, Harvard University Press, 302–33.

Soames, S. (1984). 'What is a theory of truth?', *Journal of Philosophy* 81: 411–29.

Tarski, A. (1944). 'The semantic conception of truth and the foundations of semantics', *Philosophy and Phenomenological Research* 4: 341–75.

—— (1953). 'A general method in proofs of undecidability', in *Undecidable Theories*. Amsterdam, North-Holland Publishing, 1–35.

—— (1958a). 'The concept of truth in formalized languages', in Tarski 1958c, 152–278.

—— (1958b). 'The establishment of scientific semantics', in Tarski 1958c, 401–8.

—— (1958c). *Logic, Semantics, and Metamathematics*, Corcoran, J., ed. Indianapolis, Hackett.

—— (1958d). 'On definable sets of real numbers', in Tarski 1958c, 110–42.

—— (1958e). 'Some methodological investigations on the definability of concepts', in Tarski 1958c, 279–95.