Tarski's Theory of Truth

During the 1920s and early 1930s, scientifically minded philosophers (in particular, the positivists of the Vienna Circle) regarded the notion of truth with considerable suspicion, not only on account of the liar paradox, but also because the quasi-mystical connection between language and the world in virtue of which true statements are true looked like the sort of thing properly empirical-minded philosophers ought to avoid. Alfred Tarski[1] sought to dispel these worries by showing that the notion of truth could be defined in terms of other notations whose scientific respectability was unquestioned. Let me qualify that. Tarski didn't define a general notion of truth, but rather he showed how, for a large class of languages $\mathfrak{L}$, one could define a notion of truth in $\mathfrak{L}$, applicable to the sentences of $\mathfrak{L}$. The languages involved were all formalized languages. As we shall see, the same methods cannot be used to define a notion of truth applicable to a natural language. Here we shall illustrate Tarski's methods by defining truth in the language of arithmetic.

A puzzle arises at the outset. Our current understanding of the notion of truth is insufficiently clear and precise, so we'd like to clarify the notion by providing a definition in terms that we already fully understand. But unless we already fully understand the notion of truth, how are we going to know whether the proposed definition actually succeeds? We're looking for a corrrect definition of truth, but unless we already understand the notion of truth, how will we know when we've found one?

[1] "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Logica* 1: 261-405. English translation by J. H. Woodger in Tarski *Logic, Semantics, Metamathematics*, 2nd ed. (Indianapolis: Hackett, 1983), pp. 152-278.

We faced a similar problem when we tried to answer the question, "When is a partial function computable by an algorithm?" We don't already have a precise standard of calculability – that's what we're looking for – but our intuitive understanding of the notion is sharp enough to provide us a large supply of clear examples. However, the examples we have are all one-sided: There are plenty of examples of functions that are known to be calculable, and there are examples of functions that are not known to be calculable, but we don't have any clear examples of functions that are known not to be calculable. (The Halting Problem -- determining whether a given algorithm will yield an output for a given input -- is an example of a problem that isn't solvable by algorithm, but what we call the Halting Problem is really only the sketch of a problem that we can only make precise after we have a precise characterization of calculability.) Our proposed answer, that the calculable partial functions are the Σ partial functions, met the following conditions:

> Every partial function that is calculable by known methods satisfies the criterion.
>
> Every partial function that satisfies the criterion is calculable by known methods.

This leaves open the possibility of discovering some hitherto unsuspected computational techniques that will calculate some function that doesn't meet the criterion. One can amass a fair amount of evidence to give us confidence that this won't occur, but the evidence doesn't amount to anything like a proof.

How does the situation look for trying to define truth? There are a lot of sentences that are known to be true, so that, for example, any proposed definition will have to put Fermat's Last Theorem into the extension of "true." There are other sentences that are known to like outside the extension of "true"; the negation of the Fermat theorem, for one. Also, we know that all nonsentences lie outside the extension of "true." But there are plenty of sentences for which we don't know on which side of the divide the sentence ought to fall. For example, no one now knows whether Goldbach's Conjecture ought to fall into the extension of "true" or outside it. A satisfactory definition of "true" will have to adjudicate all the unknown cases correctly, but, lacking arithmetical omniscience, how can we know when we've done this?

To be able to recognize a proposed definition of truth as satisfactory, it is not required that we know already which sentences are true. It is enough that we be able to specify, for each sentence, conditions under which it is true. Thus Goldbach's conjecture should be put into the extension of "true" if every even number greater that 2 is the sum of two primes, whereas it should be left outside the extension of "true" if not every even number is the sum of two primes. Without knowing which of these two cases obtains, we can nonetheless be sure that:

> Goldbach's Conjecture is true if and only if every even number greater than two is the sum of two primes.

This observation can be extended to a general criterion of correctness of proposed definition of truth:

> **Convention T.** A proposed definition of truth for the language of arithmetic is *adequate* if it implies all sentences of the form:
>
> (T) $\ulcorner\phi\urcorner$ is true if and only if ϕ,
>
> for ϕ a sentence of the language of arithmetic and $\ulcorner\phi\urcorner$ is Gödel number, and it also implies "$(\forall x)(x$ is true $\rightarrow$ x is a sentence$)$."

("Implies" here doesn't mean strictly logical implication. It's permissible to employ basic laws of syntax in deriving the (T)-sentences.)

With Convention T in place, it is possible actually to prove that a proposed definition of truth is adequate. You can make a strong case for the Church-Turing thesis, but you can't prove it mathematically. For truth, you can provide an honest-to-gosh proof.

Tarski's theory of truth accomplished three main things. First, he propounded the condition of adequacy, Convention T. Second, he established a range of circumstances in which the condition could be met. Third, he established a range of conditions under which the condition cannot be met.

Let's turn to the second component. Tarski wanted to provide an explicit definition of truth in nonsemantic terms, that is, he wanted to give a biconditional of the form

$(\forall x)(x \text{ is true} \leftrightarrow \tau(x))$,

where $\tau(x)$ doesn't contain any semantic terms, that is adequate in the sense of Convention T. For the language of arithmetic, here it is; keep in mind that the function Den that takes a closed term to the number it denotes can be explicitly defined by a Σ formula, so that the arithmetical expression we abbreviate Den(x) isn't really a semantic term; also the substitution operation:

$(\forall x)[x \text{ is true} \leftrightarrow$

$(\exists \text{ set } S \text{ of sentences of the language of arithmetic})$

$[(\forall \text{ closed term } u)(\forall \text{ closed term } v)(\text{Triple}(8,u,v) \in S \leftrightarrow \text{Den}(u) = \text{Den}(v)) \wedge$

$(\forall \text{ closed term } u)(\forall \text{ closed term } v)(\text{Triple}(9,u,v) \in S \leftrightarrow \text{Den}(u) < \text{Den}(v)) \wedge$

$(\forall \text{ sentence } y)(\text{Pair}(10,y) \in S \leftrightarrow y \notin S) \wedge$

$(\forall \text{ sentence } y)(\forall \text{ sentence } z)(\text{Triple}(11,y,z) \in S \leftrightarrow (y \in S \vee z \in S)) \wedge$

$(\forall \text{ sentence } y)(\forall \text{ sentence } z)(\text{Triple}(12,y,z) \in S \leftrightarrow (y \in S \wedge z \in S)) \wedge$

$(\forall \text{ sentence } y)(\forall \text{ sentence } z)(\text{Triple}(13,y,z) \in S \leftrightarrow (y \notin S \vee z \in S)) \wedge$

$(\forall \text{ sentence } y)(\forall \text{ sentence } z)(\text{Triple}(14,y,z) \in S \leftrightarrow ((y \in S \wedge z \in S) \vee (y \notin S \wedge z \notin S))) \wedge$

$(\forall \text{ formula } y)(\forall \text{ number } n)(\text{Triple}(15,n,y) \in S \leftrightarrow (\forall \text{ number } k)(y\ {}^{\mathrm{v}_n}/_{[k]} \in S) \wedge$

$(\forall \text{ formula } y)(\forall \text{ number } n)(\text{Triple}(16,n,y) \in S \leftrightarrow (\exists \text{ number } k)(y\ {}^{\mathrm{v}_n}/_{[k]} \in S) \wedge$

$x \in S]].$

Showing that this definition is adequate, as prescribed by Convention T, is labor-intensive but straightforward.

The language of arithmetic only talks about numbers, but the definition of truth for the language of arithmetic talks not only about numbers and also about sets. There is a sense in which truth is defined recursively: The truth conditions for a complex sentence are defined in terms of the truth conditions for simpler sentences. However, we cannot apply our usual technique for converting recursive definitions into explicit definitions. This technique depends on encoding finite sets of numbers by a single number, and the sets of numbers we have to talk about in defining truth are infinite. There is, in fact, no way to define the set of true sentences of the language of arithmetic within the language of arithmetic. Any definition of truth for the language of arithmetic that is adequate in the sense of Convention T must be formulated in a language richer in expressive power than the language of arithmetic. This is the third part of Tarski's theory of truth, the negative part.

To prove Tarski's result, suppose, for *reduction ad absurdum*, that there is a formula $\tau(x)$ of the language of arithmetic such that the definition

$$(\forall x)(x \text{ is true} \leftrightarrow \tau(x))$$

is adequate in the sense of Convention T. Because it's adequate in the sense of Convention T, this definition implies

$$([\ulcorner\phi\urcorner] \text{ is true} \leftrightarrow \phi),$$

for each sentence ϕ fo the language of arithmetic. In particular, we can use the Gödel self-referential lemma to find a sentence λ of the language of arithmetic such that the biconditional

$$(\lambda \leftrightarrow \neg\tau([\ulcorner\lambda\urcorner])$$

is a theorem of Q. We have the (T)-sentence

$$([\ulcorner\lambda\urcorner] \text{ is true} \leftrightarrow \lambda),$$

and the definition of "true" give us this:

$$([\ulcorner\lambda\urcorner] \text{ is true} \leftrightarrow \tau(\ulcorner\lambda\urcorner)).$$

This gives us our contradiction.⊠

We've been working with the language of arithmetic, but the same considerations apply to other languages. The general version of Convention T requires that an adequate definition of truth for a given object language entail all biconditional obtained from the schema

(T) _______________ is true if and only if ________________,

by filling in the blanks in appropriate ways. The first blank is completed with what Tarski calls a "structural-descriptive name" of a sentence of the object language. Structural-descriptive names have the property that, if you're given the name of the sentence, you can recover the sentence itself. Gödel numbers count as structural-descriptive names, as do quotation names. "The first sentence below the fold on the left column on the front page of the *Boston Globe* for May 5, 2002," does not. The second blank is filled in with the sentence's translation into the metalanguage. Tarski doesn't say anything about what makes an acceptable translation, which is a pity.

Tarski's technique for constructing an explicit definition of truth can be adapted to a wide variety of formal language. The only thing special about the language of arithmetic is that, in it, each of the individuals we are talking about is names by some numeral. In situations in which not every individual has a name, we can't define truth directly. We have to define truth in terms of satisfaction, in a way familiar from Logic I.

Tarski's negative result also generalizes, leading us to the conclusion that a definition of truth for a given object language can never be given within the language itself, but only within a richer metalanguage. For formal languages, at least most of them, that's fine; we can give the definition in plain English. The problem comes when we try to provide a definition of truth for a natural language. How can we give a definition of truth for English, when we don't have nay metalanguage richer than English?

A natural response would be to say that we don't need a definition of truth for English. We understand the notion of truth, as it applies to English, well enough without an explicit definition. Unfortunately, that reply doesn't shelter us for very long. We may not need an explicit definition of truth, but if we want to understand how language works, we at least need a theory of truth. If we look at the proof of Tarski's theory on the undefinability of truth, it doesn't show us merely that truth is undefinable. It shows us that no consistent theory of truth for a language that's formulable within that very language implies the (T)-sentences; indeed, no such theory is even consistent with the(T)-sentences.

Tarski's undefinability theorem is really just the ancient paradox of the liar, dressed up in formal wear. The paradox first appeared when Epimenides the Cretan said that Cretans always lie. A more direct version is given by Eubulides, who said "This statement is false." There is an easy response to Eubulides' paradox: declare that Eubulides' statement is neither true nor false. Easy, but short lived, for the paradox reappears if Eubulides says, "This statement is not true."

Similar paradoxes afflict other semantic notions, for example, the notion of denotation is troubled by *Berry's paradox*.[2] There are only finitely many expressions of English of fewer than thirty syllables, so, in particular, there are only finitely many expressions of English of fewer than thirty syllables that happen to name natural numbers. Each expression names at most one number, so that are only finitely many natural numbers that are named by English expressions of fewer than thirty syllables. There are, however, infinitely many natural numbers, so there are natural numbers that aren't named by any English expression of fewer than thirty syllables. Because the natural numbers are well-ordered, we know that there has to be a least natural number not named by any English expression of fewer than thirty syllables. "The least natural number not named by any English expression of fewer than thirty syllables" names it is twenty-seven syllables.

The simplest of the semantic paradoxes is *Grelling's paradox,*[3] which involves the notion of satisfaction. Some English phrases satisfy themselves and others do not. "Noun," for example, is a noun, so it satisfies itself. "Verb" isn't a verb, so it doesn't satisfy itself. "Polysyllabic," being polysyllabic, satisfies itself, whereas "monosyllabic," not being monosyllabic, doesn't satisfy itself. How about "does not satisfy itself"? Does it satisfy itself or not? Either answer leads to a contradiction.

[2] See Chapter II of Alfred North Whitehead and Bertrand Russell, *Principia Mathematica*, 2nd ed. (Cambridge: Cambridge University Press, 1927).

[3] This is a slightly different version of the paradox than the one we discussed earlier, and it's closer to the original. See Kurt Grelling and Leonard Nelson, "Bemerkungen zu den Paradoxien von Russell und Burali-Forti," *Abhandlungen der Fries'schen Schule neue Folge* 2 (1908): 301-34.

Knowledge isn't a semantic notion, but knowledge entails truth, and truth is a semantic notion. This indirect connection is enough to implicate the notion of knowledge in semantic paradox. To see this, let the *Unknown Sentence* be the following sentence:

> What the Unknown Sentence says is not known.

What the Unknown Sentence says is that what the Unknown Sentence says is not known.[4] Consequently,

> If it is not known that what the Unknown Sentence says is not known, then what the Unknown Sentence says is not known.

The first principle of epistemology – if it is known that ϕ then ϕ – gives us this:

> If it is known that what the Unknown Sentence says is not known, then what the Unknown Sentence says is not known.

Putting these two observations together, by means of an inference of the form

> If not-ψ, then θ.
>
> If ψ, then θ.
>
> Therefore, θ.

we get this:

> What the Unknown Sentence says is not known.

[4] A rather desperate attempt to evade the paradox would be to say that the Unknown Sentence doesn't express a proposition at all. It's a rather short-lived attempt, as we can see by examining the Other Unknown Sentence:

> Either the Other Unknown Sentence doesn't express a proposition, or the proposition it expresses is not known.

Now we reflect that we have reached this result by careful, explicit deduction from securely known premisses, and that things we can derive this way are known. That is,

> It is known that what the Unknown Sentence says is not known.

In other words,

> What the Unknown Sentence says is known.

Contradiction. This informal derivation can be formalized in much the way that Tarski formalized Eubulides' paradox.[5] A similar derivation can be carried out with necessity in place of knowledge.[6]

One possible response to these paradoxes, endorsed, in slightly different forms, both by Whitehead and Russell and by Tarski, is to divide us the ordinary notions of truth, satisfaction, knowledge, and so on, into infinitely many notions. Let English$_0$ be the fragment of English obtained by excising all semantic terms, together with such semi-semantic terms as "knows" and "necessary." (Exactly which terms these are isn't going to be obvious). Introduce notions of truth$_0$, falsity$_0$, satisfaction$_0$, knowledge$_0$, and so on, like the familiar notions of truth, falsity, satisfaction, and knowledge except that they're applicable only to English$_0$. For truth$_0$, this can be done simply by taking all the (T)-sentences for English$_0$ (with "true$_0$" in place of "true") as axioms. The language we get from English$_0$ by adding the predicates "true$_0$," "false$_0$," "satisfies$_0$,"

[5] See Richard Montague and David Kaplan, "A Paradox Regained," *Notre Dame Journal of Formal Logic* 1 (1960): 79-90. Reprinted in Montague, *Formal Philosophy* (New Haven: Yale University Press, 1974), pp. 271-85.

[6] See Richard Montague, "Syntactic Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability," *Act Philosophical Fennica* 16 (1963): 153-67. Reprinted in *Formal Philosophy*, pp. 286-302

"knows$_0$,"and so on, is called "English$_1$." Introduce predicates "true$_1$," "false$_1$," "satisfies$_1$," "knows$_1$," and so on applicable to the sentences of English$_1$. "True$_1$" can be introduced simply by taking the (T)-sentences for English$_1$ as axioms. Call the language obtained from English$_1$ by introducing these new predicates "English$_2$" and introduce predicates "true$_2$,""false$_2$," "satisfies$_2$," "knows$_2$," and so on, applicable to the sentences of English$_2$. This process generates a language English$_\infty$, which contains all the "true$_n$"s, "false$_n$"s, "satisfies$_n$"s and "knows$_n$"s as predicates. English$_\infty$ doesn't have global notions of truth and falsity, applicable to the whole language, but every sentence of English$_\infty$ is either true$_n$, for sufficiently large n, or false$_n$, for sufficiently large n.

This construction doesn't give a very satisfactory semantic theory for English$_\infty$, because it leaves out something that's intuitively perfectly obvious, namely, that the true sentences of English$_\infty$ are those that are true$_n$ for large enough n. And it doesn't give us a theory of truth for English at all, because English contains the predicate "true"; it doesn't contain the predicates "true$_0$," "true$_1$," "true$_2$," "true$_3$,"....

The liar paradox leaves us in a very unhappy position. As Tarski recognized, the paradox shows that one can't develop a semantic theory for a language within the language itself, but only within a richer metalanguage. But that means we lack the means to develop a semantic theory for a natural language. And without such a theory, how can we understand how human language is connected to the world around us?