Tarski's theory of truth

Jeff Speaks

March 2, 2005

We've now seen some reasons why philosophers at the time Tarski wrote were skeptical about the idea of truth. Tarski's project was, in part, to rehabilitate the notion of truth by defining the predicate 'is true' in a clear way which made use of no further problematic concepts.

1 The constraints on a definition of truth (§§1-4)

Tarski begins by saying what he thinks a definition of truth must achieve. He focuses on two constraints: material adequacy, and formal correctness.

1.1 Formal correctness

The constraint of formal correctness is the constraint that a definition of truth be given in the form,

S is true $\equiv_{df}$ _____

where the right hand side is filled in by expressions which are both free from obscurity, and do not presuppose the notion of truth to be defined.

1.2 Material adequacy

Tarski's more important and original constraint is the constraint of material adequacy. The way that he thought of this constraint shaped the form that his theory of truth eventually took.

The key passages here are in §4:

> "Consider the sentence '*snow is white.*' We ask the question under what conditions this sentence is true or false. It seems clear that if we base ourselves on the classical conception of truth, we shall say that the sentence is true if snow is white, and that it is false if snow is not white. Thus, if the definition of truth is to conform to our conception, it must imply the following equivalence:
>
> *The sentence 'snow is white' is true if, and only if, snow is white.*

Tarski thought of this as the core of the traditional conception of truth. The requirement of material adequacy is the requirement that a definition of truth conform to this traditional conception by implying each sentence of the above form. He continues:

> "We shall now generalize the procedure which we have applied above. Let us consider an arbitrary sentence; we shall replace it by the letter 'p.' We form the name of this sentence and we replace it by another letter, say 'X.' We ask now what is the logical relation between the two sentences 'X is true' and 'p.' It is clear that from the point of view of our basic conception of truth these sentences are equivalent. In other words, the following equivalence holds:
>
> (T) *X is true if, and only if, p.*
>
> We shall call any such equivalence (with 'p' replaced by any sentence of the language to which the word '*true*' refers, and 'X' replaced by a name of this sentence) and "*equivalence of the form* (T)."
>
> Now at last we are able to put into a precise form the conditions under which we will consider the usage and definition of the term '*true*' as adequate from the material point of view: we wish to use the term '*true*' in such a way that all equivalences of the form (T) can be asserted, and *we shall call a definition of truth 'adequate' if all these equivalences follow from it.*"

1.3 Other constraints

There are other constraints that one might put on a definition of truth which Tarski does not explicitly discuss. For example, on might want the definition to shed some light on various important truth-related phenomena, such as logical consequence. We will return to the relationship between a theory of truth and a theory of logical consequence later.

There are also some other obvious constraints which Tarski leaves implicit. It seems, e.g., that a theory should not entail false sentences as well as instances of the form T, and must avoid the Liar paradox. We return to the liar paradox below.

2 Tarski's definition of truth

2.1 Truth in a language

As we have seen, Tarski was interested in defining truth as a property of *sentences*. But any definition of truth for sentences must be relativized to languages. It may be one thing for a certain string of characters to be true in English, and quite another for them to be true in Italian. This means that the notion Tarski is really out to analyze is not truth *simpliciter*, but rather 'truth in L.' The task is to give a general way, for any language L, of arriving at a definition of truth for that language.

2.2 A failed definition of truth: quantifying into quotation

One might think that an obvious solution to the problem of there being an infinite number of sentences is provided by the following kind of definition of truth, in which we quantify over sentences:

x is true in $L \equiv_{df}$ for some S, x='S' & S

Why, for any of these formulae to make sense, 'x' and S' must be understood as different kinds of variables. The distinction between objectual and substitutional quantification.

Tarski gave the following objection to this kind of definition of truth in his "The Concept of Truth in Formalized Languages," which can be found in his *Logic, Semantics, and Metamathematics*:

> "Quotation-names may be treated like single words of the language, and this like syntactically simple expressions. The single constituents of these names — the quotation marks and the expressions standing between them — fulfill the same function as the letters and complexes of successive letters in single words. ... [Therefore] in applying the rule of substitution we are not justified in substituting anything at all for [a] letter ... which occurs as a component of a quotation-mark name (just as we are not permitted to substitute anything for the letter 't' in the word 'true')." (159-160)

Given this, consider the connection between the following two formulae:

x is true in $L \equiv_{df}$ for some S, x='S' & S
x is true in $L \equiv_{df}$ for some S, x=the nineteenth letter of the alphabet & S

A response to Tarski's objection to quantifying into quotation. An argument that quote-names cannot be regarded as semantically simple, since that would lead to the language's containing infinitely many semantically simple expressions.

The question of whether substitutional quantification presupposes facts about truth (and hence cannot legitimately be used in a definition of truth).

(For a classic discussion of substitutional quantification, see Kripke's "Is there a problem about substitutional quantification?" in Evans & McDowell (eds.), *Truth and Meaning: Essays in Semantics*.)

A similar problem for a structurally similar analysis of truth as a property of propositions.

2.3 Truth and satisfaction (§§5-6, 11)

Tarski calls his approach the 'semantic conception' of truth. He says in §5,

> "... the word '*true*' ... expresses a property of ... of sentences. However, it is easily seen that all the formulations which were given earlier and aimed to explain the meaning of this word ... referred not only to sentences themselves, but also to objects 'talked about' by these sentences, or possibly to 'states of affairs' described by them. And, moreover, it turns out that the simplest and the most natural way of obtaining an exact definition of truth is one which involves the use of other semantic notions, e.g. the notion of satisfaction. It is for these reasons that we count the concept of truth which is discussed here among the concepts of semantics, and the problem of defining truth proves to be closely related to the more general problem of setting up the foundations of theoretical semantics."

The next question is: how can we define truth in terms of these semantic notions? Tarski gives us an introductory sketch in §11, which is very compressed. We can sketch what he is after by explaining how to construct a Tarski-like theory of truth for languages of increasing complexity in terms of the semantic notions of designation and satisfaction.

2.3.1 A finite language: a theory of truth as a list

Consider first a very simple language, which contains only finitely many sentences. (Imagine, e.g., that the language consists only of a finite number of names and a finite number of one-place predicates.)

How could we construct a formally correct and materially adequate definition of truth for such a language? Well, we need something of the form given above, which entails each of the instances of form (T) for the language. If the language contains, say, 100 sentences, then the following seems to do the trick:

$$x \text{ is true in } L \equiv_{df} [(S_1 \ \& \ x = \text{“}S_1\text{”}) \text{ or } (S_2 \ \& \ x = \text{“}S_2\text{”}) \text{ or } \ldots \text{or } (S_{100} \ \& \ x = \text{“}S_{100}\text{”})]$$

This is of the right form, and implies all of the instances of schema (T). To that extent, Tarski is willing to regard this list-like definition as a perfectly satisfactory definition of truth in this finite language. The problem with this definition in Tarski's view is not that it is incorrect, but that it is insufficiently general: this recipe for giving theories of truth fails to apply to languages, like English and any other natural language, which contain an infinite number of sentences.

Another worry about list-like definitions: they can't be used in definitions of logical consequence. (Though, in a finite language, one could also give a list-like definition of logical consequence.)

2.3.2 A simple language with infinitely many sentences

Suppose first that we have a language like our simple 100-sentence language above, but with the addition of two new pieces of vocabulary: '¬' and '&'. Now we have infinitely many sentences in the language, and so cannot give a list-style definition of truth in the language.

For this kind of language, we can give an intuitive inductive definition of truth with the following axioms:

An atomic sentence x is true in $L \equiv_{df}$ $[(S_1$ & x="S_1") or (S_2 & x="'S_2") or ... or (S_{100} & x="S_{100}")]

A sentence of the form ⌜¬x⌝ is true in $L \equiv_{df}$ x is not true

A sentence of the form ⌜x & y⌝ is true in $L \equiv_{df}$ x is true and y is true

How this can be converted into an 'explicit definition' of truth which is formally correct, in Tarski's sense.

2.3.3 A language with function symbols

Though this language is more complex than our original finite language, it is still a very simple language. One respect in which it is simplified is that, if we think of the atomic sentences of the above language as the ones with no occurrences of '¬' or '&', it has only finitely many atomic sentences. But natural languages, for example, are not like this.

Suppose we introduce into the language some functional expressions, which can be combined with a singular term to form a new singular term. An intuitive example might be the expression 'The mother of'. This can be combined with a singular term to form a new singular term; moreover, it can be iterated indefinitely many times, giving us indefinitely many singular terms in the language. For such a language, the definition of truth given above will obviously be unsatisfactory, since the above definition relied on a list-like definition of truth for atomic sentences. But now we have an indefinite number of atomic sentences.

To make any progress here, we will obviously have to give some kind of definition of truth for atomic sentences which goes beyond the list-like definitions given above. Let's suppose for simplicity that every atomic sentence consists of one singular term combined with a one-place predicate. Then it looks like we can give following theory of truth for atomic sentences, which can be used to replaced the list-like definition above:

An atomic sentence ⌜n is F⌝ is true in L iff the object designated by ⌜n⌝ satisfies ⌜is F⌝.

But you might have the following objection: even if this seems true so far as it goes, it hardly satisfies Tarski's 'formal correctness' constraint. After all, this uses the undefined semantic notions of designation and satisfaction in the analysis of truth.

2.3.4 Definitions of designation and satisfaction

How can we analyze designation and satisfaction?

The answer to this question is in some ways reminiscent to our answer of how to construct a theory of truth for a language with only finitely many sentences. So see how, first suppose that our language has only three names and three predicates, 'Bob', 'Jane', and 'Nancy' and 'is nice', 'is mean,' and 'is lazy.' We can then give the following analyses of designation and satisfaction for the language:

> Definition of designation
>
> a name n designates an object $o \equiv_{df}$ [(n="Bob" and o = Bob) $\vee$ (n="Jane" and o = Jane) $\vee$ (n="Nancy" and o = Nancy)]
>
> Definition of satisfaction
>
> an object o satisfies a predicate $p \equiv_{df}$ [(p="is nice" and o is nice) $\vee$ (n="is mean" and o is mean) $\vee$ (n="is lazy" and o is lazy)]

You should notice that there is an analogy between the material adequacy constraint which Tarski set on the theory of truth, and similar constraints which we should expect our definitions of designation and satisfaction to meet. Just as a theory of truth for a language should imply every instance of

> 'S' is true in L iff S

so we should expect our theories of designation and satisfaction to imply every instance of the following two schemata:

> 'n' designates o in L iff $o = n$
>
> o satisfies 'is F' in L iff o is F

Here you might object as follows: the list-like definition of truth failed for a language with infinitely many sentences. But now, with the introduction of functional expressions like 'The mother of', we have infinitely many singular terms. So how can we give a list-like definition of designation?

The answer is that we cannot, and that the above definition of designation is only satisfactory for a language which contains, as its singular terms, only a finite stock of simple names. Suppose that in our language we have, in addition to the three names above, two functional expressions: 'The mother of' and 'The father of.' Since these can be iterated indefinitely, this gives us infinitely many singular terms in the language.

Then we can give the following two part definition of designation:

> A name n designates an object $o \equiv_{df}$ [(n="Bob" and o = Bob) $\vee$ (n="Jane" and o = Jane) $\vee$ (n="Nancy" and o = Nancy)]
>
> A functional expression $\ulcorner f(s) \urcorner$ designates an object $o \equiv_{df}$ [($\ulcorner$f$\urcorner$='The mother of' & for some object o', $\ulcorner$s$\urcorner$ designates o' & o is the mother of o') $\vee$ ($\ulcorner$f$\urcorner$='The father of' & for some object o', $\ulcorner$s$\urcorner$ designates o' & o is the father of o')]

This entails every instance of the 'designation schema' given above, for each of the infinitely many singular terms of the language.

(The treatment of complex predicates runs in an analogous way. See the discussion of "x is greater than y or x is equal to y" in §11, p. 353.)

2.3.5 A language with quantifiers

One important respect in which the above language is simplified is that every non-atomic sentence in the language is a combination of two or more whole sentences. This was clear from the way we proceeded: we first gave definitions of truth for atomic sentences, and then explained the truth of non-atomic sentences in terms of the truth of the atomic sentences.

But if we add quantifiers to the language, we get sentences which are neither atomic nor truth-functions of whole sentences, like

> $\exists x$ red (x)

So let's consider a language like the one we discussed above, but with the addition of the existential quantifier. How can we handle quantified sentences in our definition of truth?

Let's assume, to simplify, that every object in the domain has a name. Then, given this assumption, it looks like the intuitive thing to say about the sentence above is that it is true just in case there is some name such that, if we replaced the variable with the name and deleted the quantifier, we would get a true sentence. This, it seems, provides a definition of the truth of existentially quantified sentences in terms of our definition of truth for atomic sentences.

This is useful, but it is not in general true that languages contain names for every object in the domain.

Assignments of values to variables, and Tarskian definitions of truth relative to an assignment.

2.4 The liar and the hierarchy of languages (§§7-10)

Th foregoing provides some explanation of how to construct Tarskian truth theories for languages of increasing complexity. But it provides no explanation, you might think, for how the Liar paradox is to be avoided.

It is clear that Tarski takes the Liar Paradox to be a very serious challenge to any attempt to make sense of the concept of truth. About the Paradox, he says,

> "In m y judgement, it would be quite wrong and dangerous from the standpoint of scientific progress to depreciate the importance of this and other antinomies, and to treat them as jokes or sophistries. It is a fact that we are here in the presence of an absurdity, that we have been compelled to assert a false sentence. ... If we take our work seriously, we cannot be reconciled to this fact. We must discover its cause, that is to say, we must analyze premises upon which the antinomy is based; we must then reject at least one of these premises, and we must investigate the consequences which this has for the whole domain of our research." (348)

Tarski goes on (in §8) to state the three assumptions about the sample language L which lead to "the antinomy of the Liar":

1. L contains the resources for stating facts about its own semantics, such as the term 'true' applying to sentences of L. Tarski calls this L being 'semantically closed.'

2. The ordinary laws of logic hold, including the law of the excluded middle.

3. The language contains the capacity to refer to its own expressions.

Since these three premises lead to a contradiction, we must reject one of them. The problem is that each of the three seem very plausible. Tarski thinks that it is clear that we cannot reject the second assumption, and it is extremely implausible to reject the third. So Tarski concludes that we must reject the first of the three assumptions: "Accordingly, we decide *not to use any language which is semantically closed* in the sense given." (349) According to Tarski, no language can state its own semantics.

This raises the question of what we are doing when we state a Tarskian theory of truth for a language. Are we not contravening the decision not to use any language which is semantically closed? Tarski turns to this question in §9.

His answer is that we must distinguish the language for which we are giving a theory of truth, and the language in which we state the theory. He calls the former the 'object-language', and the latter the 'meta-language.' He concludes: "In this way we arrive at a whole hierarchy of languages." (350)

The requirements on the meta-language. Why everything statable in the object-language must be statable in the meta-language.

How the rejection of assumption (1) provides a solution to the Liar Paradox.

3 The importance of Tarski's definition

3.1 The reply to truth skepticism

Tarski's definition as a definition of truth, designation, and satisfaction for a language which makes no use of concepts other than those employed in the language itself. Hence, if the original language was legitimate, so is that language with a Tarskian truth predicate added to it. The proof that a truth predicate with these characteristics could be rigorously defined for some formalized languages was one of Tarski's main achievements.

3.2 The definition of logical consequence

How Tarski's definition of truth can be used to define logical consequence (for a language).

4 Natural and formalized languages

In §6, Tarski says that his definition only applies to truth predicates in formalized languages, and not to natural languages:

> "For other languages — thus, for all natural, 'spoken' languages — the meaning of the problem is more or less vague, and its solution can have only an approximate character. Roughly speaking, the approximation consists in replacing a natural

> language ... by one whose structure is exactly specified, and which diverges from the given language 'as little as possible'."

But, you might wonder, can Tarski's account explain the meaning of the word "true", as we use it in English?

It seems clear that Tarski is not very interested in this question. He says in §14:

> "I hope nothing which is said here will be interpreted as a claim that the semantic conception of truth is the 'right' ... one. I do not have the slightest intention to contribute in any way to those endless, often violent discussions on the subject ... I must confess I do not understand what is at stake in such disputes. ...
>
> It seems to me obvious that the only rational approach to such problems would be the following: We should reconcile ourselves with the fact that we are confronted, not with one concept, but with several different concepts which are denoted by one word; we should try to make these concepts as clear as possible ... to avoid further confusions, we should agree to use different terms for different concepts; and then we may proceed to a quiet and systematic study of all the concepts involved ..."

Why this response is not entirely satisfactory.

We can raise at least three problems for the attempt to apply Tarski's theory of truth to a natural language like English.

4.1 Natural languages lack appropriate formalizations

The case of propositional attitude ascriptions; the problem of giving a truth theory for sentences like

> Galileo believed that the earth is round.

Why Tarski was not interested in sentences like this one. For an attempt to expand Tarski's model to handle sentences like this one, see Davidson, "On Saying That."

4.2 'True' in English applies primarily to propositions, not sentences

4.3 Natural languages seem to be semantically closed

Tarski's undefinability theorem: the argument, based on the Liar, that no bivalent language with the capacity for self-reference and some basic logical resources can contain its own truth predicate. However, English, e.g., may well seem to be such a language.

Tarski's response to the problem of the semantic closure of natural languages:

> "A characteristic feature of colloquial language ... is its universality. If we are to maintain this universality of everyday language in connexion with semantical investigations, we must, to be consistent, admit into the language, in addition to its

> sentences and other expressions, also the names of these sentences and expressions, and sentences containing these name, as well as such semantic expressions as 'true sentence', 'name', 'denote', etc. ... But it is presumably just this universality of everyday languages which is the primary source of all semantical antinomies, like ... the liar ... These antinomies seem to provide a proof that every language which is universal in the above sense, and for which the normal laws of logic hold, must be inconsistent." (from "The Concept of Truth in Formalized Languages")

Tarski's view that natural languages are inconsistent. A problem with the interpretation of this claim.

A different response to the problem that natural languages seem to be semantically closed: natural languages as containing a (unsubscripted) hierarchy of truth predicates.

...

For further helpful readings see Soames, *Understanding Truth*, and the entry on Tarski's definition of truth by Anil Gupta in the Routledge Encyclopedia of Philosophy.