Lecture 1: Tarski on Truth

Philosophy of Logic and Language — HT 2016-17

Jonny McIntosh
jonathon.mcintosh@balliol.ox.ac.uk

Alfred Tarski (1901-1983) was a Polish (and later, American) mathematician, logician, and philosopher. In the 1930s, he published two classic papers: 'The Concept of Truth in Formalized Languages' (1933) and 'On the Concept of Logical Consequence' (1936). He gives a definition of truth for formal languages of logic and mathematics in the first paper, and the essentials of the model-theoretic definition of logical consequence in the second. Over the course of the next few lectures, we'll look at each of these in turn.

1 Background

The notion of truth seems to lie at the centre of a range of other notions that are central to theorising about the formal languages of logic and mathematics: e.g. validity, consistency, and completeness. But the notion of truth seems to give rise to contradiction:

Let sentence (1) = 'sentence (1) is not true'. Then:

1. 'sentence (1) is not true' is true IFF sentence (1) is not true
2. sentence (1) = 'sentence (1) is not true'
3. So, sentence (1) is true IFF sentence (1) is not true
4. So, sentence (1) is true and sentence (1) is not true

Tarski is worried that, unless the paradox can be resolved, metatheoretical results invoking the notion of truth or other notions that depend on it will be remain suspect. But how can it be resolved? The second premise is undeniable, and the first premise is an instance of a schema that seems central to the concept of truth, namely the following:

> 'S' is true IFF S,

According to Tarski, the problem arises out of the fact that natural languages are *semantically closed*: for each sentence S, they contain another sentence S' that attributes truth

(untruth) to S. In order to establish the sorts of metatheoretical results that Tarski is interested in, however, we need to be able to talk about the true and untrue sentences of formal languages. How are we do that unless those languages are semantically closed?

Tarski's solution: talk about the true and untrue sentences of one language, called the *object language*, in another language, called the *metalanguage*. Tarski shows how, given an object language L and metalanguage M that meet certain requirements, we can define a predicate $True_L$ *in M* that applies to all and only the true sentences *of L*.

2 The Shape of the Problem

2.1 Formal Correctness

Tarski imposes various requirements on a satisfactory definition of $True_L$. First, it must be *formally correct*, i.e. it must be (or be provably equivalent to) a sentence of the form,

$\forall x, True_L(x)$ if and only if $\phi(x)$,

where ϕ doesn't contain the predicate $True_L$ or any other predicates expressing otherwise obscure notions. (Definitions taking this form are called *explicit* definitions.)

This requirement ensures that the definition is not circular. However, it's obviously not enough to ensure that what is defined is a *truth* predicate. Consider, for example:

$\forall x, True_L(x)$ if and only if $x = x$.

This is formally correct, but obviously inadequate: it counts *every* sentence of L as true.

2.2 Material Adequacy

The definition must also be *materially adequate*. This is the requirement that the definition capture the core of the meaning of the word 'true'. To do this, it must at the very least be *extensionally adequate*. That is to say, unlike the definition above, it should yield a predicate $True_L$ that applies to all and *only* the true sentences of the object language.[1]

Tarski suggests that a (formally correct) definition of $True_L$ for a given object language L is materially adequate IFF it entails, for each sentence of L, an instance of the schema:

(T) 'X' is $True_L$ IFF p,

where 'X' is replaced by a name of the sentence and 'p' by a translation of that sentence in M. (If M is just an extension of L, 'p' will be replaced by the sentence itself.)

[1] It is controversial whether Tarski intended or achieved anything more than extensional adequacy.

Tarski's suggestion here is his famous *Convention T*, though a better name for it would be *Criterion T*. To see the idea, suppose teenagers start using the predicate 'peng'.

On the one hand, it seems that 'peng' can only be a truth predicate for English if each instance of the following schema, where what replaces 'X' is a name of an English sentence and what replaces 'p' is either that sentence or a paraphrase of it, is true:

(P) 'X' is peng IFF p,

For example, it will have to be the case that 'snow is white' is peng IFF snow is white.

This seems to show that it is a *necessary* condition on 'peng' being a truth predicate for English that each relevant instance of schema **(P)** is true. On the other hand, it also seems to be a *sufficient* condition. Put roughly, the thought is this. Suppose each relevant instance of **(P)** is true. This tells us that, if a sentence of English says that p, then that sentence is peng IFF p. Yet it also seems that, if a sentence of English says that p, it is a true sentence of English IFF p. It thus follows that it is peng IFF it is a true.

2.3 Tarski's Hierarchy

Various constraints on the relationship between L and M fall out of Convention T:

- M has to contain resources for *referring* to each of the sentences of L.
- M has to contain the sentences of L (or their *translations*).

Convention T doesn't in and of itself rule out the possibility that M is identical to T. But if it is, T will be semantically closed. And we saw at the outset that Tarski thought that this leads to the Liar Paradox. So Tarski also thinks that M has to be distinct from T.

Here's a way of seeing the problem Tarski has in mind. Suppose that M is identical to T. Then $True_L$ can be defined in L itself and, so long as L contains a negation operator, there will be a sentence in L, whose name in L is λ, and which is of the form, $\neg True(\lambda)$ – i.e. a sentence that "says of itself" that it is not true. In accordance with Convention T, the definition of $True_L$ will entail a theorem of the form $True(\lambda) \equiv \neg True(\lambda)$. Assuming classical logic, the Liar Paradox follows. If we want to (a) retain classical logic and (b) allow for the possibility of languages that contain the resources for referring to their own sentences (and a negation operator), we have to reject the assumption that $M = L$.

Similarly, M cannot contain *its* own truth predicate. So to talk about the true sentences of M, we need a distinct *metametalanguage* M', giving rise to a hierarchy of languages.

If we distinguish between object languages and their metalanguages, the Liar Paradox is avoided. The truth predicate defined in the metalanguage will only apply to sentences of the object language, and so to sentences that do not contain that predicate.

3 How Tarski Solves It

3.1 Finite Languages

Tarski also shows how to construct definitions along these lines for various sorts of languages. It's very easy in some simple cases. Consider L_1, which contains just two sentences: '1 + 1 = 2' and '1 + 1 = 3'. Assuming these have their ordinary meanings, a definition of a truth predicate for L_1, using English as our metalanguage, might be:

> $\forall s$(s is true$_{L_1}$ IFF
> ((s is '1 + 1 = 2' and one plus one is two) OR
> (s is '1 + 1 = 3' and one plus one is three)))

- Is this formally correct?
- Is it also materially adequate?

3.2 Recursive Definitions

For 'true$_{L1}$', we just listed out the truth conditions of the various sentences of L_1. So long as the language contains only a finite number of sentences, we can always take this sort of approach. But what about when it doesn't contain a finite number of sentences?

Suppose, for example, we consider L_2, which extends L_1 with the sentential connectives, '$\neg$' and '$\wedge$' — where these have the meanings that you're familiar with from 1st year logic. These operators can be iterated, so L_2 contains infinitely many sentences. In this case, Tarski suggests that we give a *recursive* definition: we first define the truth predicate for a certain basic set of sentences of the language, the *atomic* sentences, and then extend the definition to all the other sentences in terms of what we have said about the basic set. For example, a recursive definition of a truth predicate for L_2 might be:

> $\forall s$(s is true$_{L_2}$ IFF
> ((s is '1 + 1 = 2' and one plus one is two) OR
> (s is '1 + 1 = 3' and one plus one is three) OR
> (s is formed by prefixing a sentence s_1 with '$\neg$' and s_1 is not true$_{L_2}$) OR
> (s is formed by placing '$\wedge$' between sentences s_1 and s_2 and both s_1 and s_2 are true$_{L_2}$)))

- Is this formally correct?

  We might worry it's not, as 'true$_{L_2}$' occurs on the RHS. But as Frege and Dedekind showed, given some mathematical machinery, it's possible to turn recursive definitions like this into explicit, eliminative definitions.

- Is it materially adequate?

3.3 Quantification

Quantifiers pose a further problem. We cannot specify the truth conditions of quantified sentences, such as '$\exists x(x + x = 2)$', in terms of the truth values of their parts, as their parts don't generally *have* truth values. Tarski offers two methods for dealing with this. The first involves replacing variables with names of the objects in the domain of quantification. But this won't always work, as in general some objects in the domain of quantification won't have names. So for the general case, Tarski developed the method that will be familiar from the semantics of predicate logic. This involves the notion of a *variable assignment*, and treats the variables themselves as a kind of *temporary* name.

4 Philosophical Significance

At the very least, Tarski shows us how to give explicit definitions of predicates that are co-extensive with various substitution instances of the predicate 'is a true sentence of L'. This in itself is an impressive achievement. But does he do more?

4.1 Defining the Concept of Truth?

Does Tarski provide the means for defining the concept of truth, i.e. for defining predicates that have the same meaning as, say, the English predicate 'is true'?

Problems:

- The predicate 'is true' applies to propositions, not sentences.

  OK. But if so, we can ask whether Tarski provides the means for defining predicates that have the same meaning as the English predicate 'expresses a truth'.

- The English predicate 'is true' (or 'expresses a truth') applies to sentences of a *range* of languages, including English! Tarski's predicates only apply to one language, and do not apply to sentences of the language to which they belong. If Tarski's diagnosis of the Liar Paradox is right, this is a serious objection: it means his predicates *cannot* be co-extensive with 'is true' or 'expresses a truth'.

4.2 Explicating the Concept of Truth?

Perhaps we could more modestly take Tarski as providing the means for *explicating* the concept of truth, i.e. for defining predicates that can replace the predicate 'is true' (or 'expresses a truth') in all legitimate theoretical contexts, but that lack of any of its defects — and in particular, don't give rise to the Liar Paradox!

Problems:

- Tarskian definitions, such as the definition of 'true_{L_2}', might be regarded in either of two ways: as (1) telling us what the predicate being defined means or (2) as

telling us what the sentences of the object language mean. The second way of regarding them depends on the idea that to understand a sentence is to know its truth conditions, an idea that lies at the heart of *truth conditional semantics*.

The problem is that it seems that a definition of truth cannot accomplish both tasks at the same time. To accomplish the second, we would have to already know what the point of describing a sentence as 'true$_{L_2}$' is. (Contrast: a theory that tells us under what conditions the sentences of a language are *peng*!) But telling us *that* is part of the first task. We can call this the *problem of truth conditional semantics*.

One way to solve the problem of truth conditional semantics is to reject the idea that to know what the sentences of a language mean is to know their truth conditions. This idea will be discussed in more detail next term's lectures for 108.

- Various philosophers have raised the following *problem of modal difference*. Consider the definition of 'true$_{L_1}$':

  $\forall s$(s is true$_{L_1}$ IFF
  ((s is '1 + 1 = 2' and one plus one is two) OR
  (s is '1 + 1 = 3' and one plus one is three)))

  Now, if we replace 'true$_{L_1}$' in

  '1 + 1 = 2' is true$_{L_1}$

  with its definiens, we get:

  ('1 + 1 = 2' is '1 + 1 = 2' and one plus one is two) OR ('1 + 1 = 2' is '1 + 1 = 3' and one plus one is three)

  But now, '1 + 1 = 2' has this property in all worlds in which one plus one is two, i.e. in all worlds. But it is surely not *true* in all worlds: for it is surely false in worlds in which '2' refers to three and all the other words have their actual meaning?

  But the languages for which Tarski's definitions are given are individuated by their semantics as well as their syntax. So it is a necessary truth that '1 + 1 = 2' means in L_1 that one plus one is two. So it is true in all worlds after all.

- Max Black first raised a *problem of non-projectability*: definitions like those of 'true$_{L_1}$' and 'true$_{L_2}$' don't *project*. They don't tell us under what conditions other truth predicates, for other languages, hold of the sentences of those languages.

- There is also a *problem of epistemic difference*. Knowing the truth conditions of a sentence, whether or not it's sufficient for knowing what that sentence means, at least gives some negative information about its meaning. But knowing the conditions under which, say, '1 + 1 = 2' is true$_{L1}$ doesn't even provide that.

Selected Bibliography

Starred items (*) are more introductory, and good places to start. Tarski (1933) is his first proper treatment of the topic, but his (1944) is a more accessible presentation. Even more accessible still is his (1969). Of the rest, Field (1972), Etchemendy (1988), Soames (1999), and Künne (2003) will be particularly useful in starting out on essays.

*Alexis P. Burgess and John P. Burgess (2010) *Truth* (Princeton UP), Ch. 2

Donald Davidson (1990) 'The Structure and Content of Truth' in *Journal of Philosophy* 87(6), pp. 279-328.

Michael Dummett (1959) 'Truth' in *Proceedings of the Aristotelian Society* 59, pp. 141-162.

John Etchemendy (1988) 'Tarski on Truth and Logical Consequence' in *The Journal of Symbolic Logic*, 53(1), pp. 51-79, §1.

Hartry Field (1972) 'Tarski's Theory of Truth' in *Journal of Philosophy*, 69(13), pp. 347-375.

*Mario Gómez-Torrente (2006) 'Alfred Tarski' in E. Zalta, ed. *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/tarski/

Richard Heck, Jr. (1997) 'Tarski, Truth, and Semantics' *The Philosophical Review* 106(4), 533-554.

*Wilfrid Hodges (2001) 'Tarski's Truth Definitions' in E. Zalta, ed. *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/tarski-truth/

*Richard Kirkham (1992) *Theories of Truth* (MIT Press), Ch. 5 and 6.

Wolfgang Künne (2003) *Conceptions of Truth* (OUP), §4.2.

Douglas Patterson (2011) *Alfred Tarski: Philosophy of Language and Logic* (Palgrave Macmillan), Ch. 4.

Hilary Putnam (1985) 'A Comparison of Something with Something Else' in *New Literary History* 17(1), pp. 61-79.

Gila Sher (1999) 'What is Tarski's Theory of Truth?' in *Topoi* 18(2), pp. 149-166.

Scott Soames (1999) *Understanding Truth* (OUP), Ch. 3 and 4.

Alfred Tarski (1933) 'The Concept of Truth in Formalized Languages' in his (1983) *Logic, Semantics, and Metamathematics*, 2nd revised edition (Hackett).

— (1944) 'The Semantic Conception of Truth' in *Philosophy and Phenomenological Research* 4(3), pp. 341-376.

— (1969) 'Truth and Proof' in *Scientific American* 220, pp. 63-77.