How Tarski defined the undefinable

CEZARY CIEŚLIŃSKI

Logic Department of the Institute of Philosophy, University of Warsaw
E-mail: c.cieslinski@uw.edu.pl

This paper describes Tarski's project of rehabilitating the notion of truth, previously considered dubious by many philosophers. The project was realized by providing a formal truth definition, which doesn't employ any problematic concept.

Alfred Tarski, one of the greatest logicians of all time, was born as Alfred Tajtelbaum in 1901. He came from a Jewish family living in Warsaw. In 1918-24 he studied at the University of Warsaw, initially biology, and then mathematics. At this time Warsaw was becoming one of the world leading centers in the area of logic and mathematics. Tarski's logic teachers were Jan Łukasiewicz and Stanisław Leśniewski, prominent members of the emerging Lvov-Warsaw school – a thriving community of researchers and students, which was to flourish in Poland until the beginning of World War II.[1] Tarski wrote his PhD dissertation "On the primitive term of logistic" under the supervision of Stanisław Leśniewski (the dissertation having to do with Prothotetics – a logical system devised by his supervisor).[2] In later years Leśniewski kept saying "All my PhD students are geniuses". It was widely known however that Tarski was the only PhD student of Leśniewski.

After completing the doctorate, Tarski became a docent at the University of Warsaw. He was at that time the youngest researcher working as a docent in Warsaw. Since his job was poorly paid, he also taught mathematics at a secondary school in Warsaw.

In August 1939 Tarski left for the US to take part in the Unity of Science Congress in Harvard. The invitation to address the Congress probably saved his life – World War II broke out soon after his departure from Europe. He remained in America for the rest of his life, teaching in Harvard and Berkeley, where he created a prominent center of research in logic and foundations of mathematics. He died in Berkeley in 1983.[3]

Tarski was a prolific scientist with a very broad scope of interests. His influence on contemporary logic is enormous; it is worth stressing however that he made important contributions not only to logic, but also to set theory, topology, geometry, and algebra. In a paper written together with Stefan Banach he proved the theorem on the decomposition of the sphere, which states that a ball can be decomposed into a finite number of pieces and then reassembled into a ball larger than

[1] For a comprehensive history of the Lvov-Warsaw school, see Woleński (1989).
[2] See Tarski (1923).
[3] Tarski's biography by Feferman, A., and Feferman, S. (2004) contains a lot of information about Tarski and his milieu, both in Poland and in the US.

the original one. This result is known nowadays as "Banach-Tarski paradox".[4] He proved the decidability of elementary theory of real numbers, which was a very surprising result given the fact that elementary theory of natural numbers was known to be undecidable. Other noteworthy achievements include a new axiomatization of geometry and the pioneering work on relation algebras.

However, for many philosophers and logicians it is Tarski's work on truth that is the highlight of his scientific contributions.[5] Also to the wider audience Tarski is known first of all as "the man who defined truth", although it should be admitted that the popular reception of his results is seriously limited by their technical, mathematical character. Accordingly, my main aim in this paper is to describe Tarski's work on truth in a non-technical manner, making it as accessible as possible to the lay reader. Initially the results obtained by Tarski may sound baffling. On the one hand, he showed how to construct a mathematically precise truth definition. On the other, his famous undefinability theorem states that the notion of truth is undefinable. How is it possible to have it both ways? How can one proclaim undefinability and define something, all in one go? In what follows I will try to answer this question, sketching the basic ideas behind Tarski's construction.

Traditional explanations

Since ancient times philosophers have been asking questions about the nature of truth. Let us start with some classical quotes, where the attempt is made to produce answers.

- "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true." Aristotle, *Metaphysics*, Γ, 7, 27.
- "Veritas est adaequatio intellectus et rei." (Truth is the conformity of the intellect to the things.) Thomas Aquinas, *Summa Theologica* I, Q 16.
- "The nominal definition of truth, namely that it is the agreement of cognition with its object, is here granted and presupposed." I. Kant, *Critique of Pure Reason*, A 57-8/B 82.

The formulations given by Aquinas and Kant are very typical in one respect: they employ a special relation, which is supposed to hold between intellect (or cognition) and its object: the one of "conformity" or "agreement". The term "correspondence" is also used in the literature with a similar intent: in short, truth is understood as a correspondence of thought (cognition) with reality. On the other hand, Aristotle's formulation seems to be much more cautious in this respect: it does not

[4] See Banach, S., and Tarski, A. (1924). Here is a popular jocular version of the paradox: a pea can be decomposed into finitely many pieces; then these pieces can be rearranged forming a ball the size of the sun. A word of caution however: in fact the theorem does not translate in such a way into physical reality!

[5] The primary source is Tarski (1933); see also Tarski (1944).

invoke any correspondence relation, the impression is rather that of an austere and minimalistic approach to the notion of truth. To say "there are goblins" is false, since there are no goblins; to say "there are horses" is true since there are horses – in contrast to Aquinas and Kant, all talk of the correspondence relation is avoided here. This minimalist spirit is quite popular in contemporary philosophy; some elements of this approach can be found also in Tarski's work.

Traditional explanations of the notion of truth encountered two basic problems. First, several troublesome questions can (and should) be asked about the correspondence relation. Second, semantic theories (including theories of truth) have been plagued by semantic paradoxes. We are going to discuss both issues in turn.

Correspondence relation

If truth is correspondence with reality, the following questions seem very natural:

(a) What sorts of objects can correspond to reality in the desired sense? In short: what objects are truth-bearers?
(b) Which fragments of reality are supposed to stand in a correspondence relation to the truth-bearers?
(c) What is the nature of the correspondence relation?

Various answers can be given to question (a). One option consists in attributing truth to *sentences* – linguistic objects characterized in purely syntactic terms. On this approach, when we call true the English sentence "Snow is white" and the German sentence "Schnee ist weiss", these are two different truth attributions, since the objects (i.e. the sentences) are clearly different in the two cases, even if they mean the same. Alternatively, one could treat not sentences, but *propositions* as primary truth-bearers. Proposition is understood as the content expressed by the sentence. Two different sentences may express the same proposition, e.g. "Snow is white" and "Schnee ist weiss" express the same proposition (namely, the proposition that snow is white) and it is propositions, not sentences, which on this view are evaluated as true or false. Still another candidates for the role of truth-bearers could be considered, for example thoughts, conceived as psychological objects. Then we could say e.g. that a particular thought of John is true (perhaps yesterday in the afternoon he had a thought that snow is white). Be that as it may, any decent explanation of the notion of truth should pick some objects for the role of truth-bearers.

As to the question (b), the issue here is whether a given truth-bearer corresponds to reality taken as a whole, or just to a fragment of it, and if it is a fragment, how can we specify which one? Indeed, both moves – holistic and fragmentary one - have been tried out by philosophers. A

fragmentary approach may consist in declaring that true sentences correspond to *facts* or *states of affairs*. For example, there is the fact that George broke his leg; accordingly, the sentence "George broke his leg" would correspond to this concrete fact. (But let us note in passing that such a solution generates only new philosophical questions about the nature of facts – what are they exactly?)

Question (c) is also troublesome. What is the nature of the correspondence relation between a given truth-bearer and reality, say: between a proposition and a fact it describes? Is some similarity of structure required – should a proposition be structurally similar to the corresponding fact? To put it in another way, in virtue of *what* does the correspondence relation hold?

At the beginning of the XX century many philosophers were deeply dissatisfied with traditional answers to these and related questions. But that was only the part of the trouble, as semantic paradoxes endangered the very consistency of our intuitive talk about truth.

Semantic paradoxes

Many semantic paradoxes have been discovered, but the most famous of them – the king of all semantic paradoxes – is the celebrated liar paradox. Consider the sentence (L) stating its own untruth. That is, let (L) be the sentence:

(L) is false.[6]

Is (L) true or false? A simple reasoning convinces us that no answer to this question is viable. We carry out the reasoning considering each of the two cases separately.

Case 1: assume that (L) is true. Then it is exactly as (L) states; but since (L) states that (L) is false, it follows that (L) is indeed false, contrary to the initial assumption that (L) is true.
Case 2: (L) is false. Then it is not as (L) states; and since (L) states that (L) is false, it follows that (L) is not false, contrary to the initial assumption.

Paradoxes of this sort have plagued truth theories for ancient times. The oldest version of the liar paradox is attributed to the Greek philosopher Eubulides of Miletus, who asked whether a man saying that he is lying is telling the truth, or lying (hence the name "liar paradox"). Another ancient version is known as "Epimenides paradox". Imagine that Epimenides, a Cretan, says that all Cretans always lie. Is he telling the truth? It is worth noting that in spite of its apparent similarity to Eubulides' liar, this version does not produce any inconsistency. Indeed, the assumption that all

[6] Alternatively, imagine a book where the first sentence on page 1 reads: "The first sentence on page 1 in this book is false". Since it is in fact the first sentence on page 1 (we can check it empirically), it states in effect its own falsity. The formulation provided above does not require such an empirical checking.

Cretans always lie leads straight to a contradiction: Epimenides is a Cretan after all and if the assumption holds, he has just told the truth! But this shows only that some Cretans sometimes tell the truth and Epimenides lied. No further contradiction follows. In this respect the reasoning differs radically from other versions of the liar, where the real contradiction is produced.

In view of these problems, at the beginning of the XX century the talk of truth was often treated as unscientific. To be sure, there were philosophers who attempted to explain the notion of truth, but no rigorous and systematic analysis was available. There was in fact much skepticism in the air. As Kurt Gödel put it:

> In consequence of the philosophical prejudices of our times […] a concept of objective mathematical truth as opposed to demonstrability was viewed with greatest suspicion and widely rejected as meaningless.[7]

Semantics – a discipline investigating languages together with their interpretations, ascribing content to linguistic expressions – was treated as pseudoscience, on a par with metaphysics. In Carnap's words:

> [Many philosophers] seem to think that pragmatics - as a theory of the use of language - is unobjectionable, along with syntax […] but semantics arouses their suspicions. They are afraid that a discussions of […] truth - as distinguished from confirmation by observations - will open the back door to speculative metaphysics, which was put out at the front door.[8]

Tarski's work had a revolutionary impact, consisting in rehabilitating the notion of truth as a respectable scientific concept, and initiating a new discipline (called later "model theory"), devoted to the study of languages and their interpretations. In what follows I will sketch the basic features of Tarski's approach.

Tarski's approach

Tarski explicitly distanced himself from any attempt to define the notion of truth for natural languages, like English. According to Tarski, the use of the predicate "true" in natural languages produces inconsistencies (as in the liar paradox). Therefore rigorous characterization of the notion of truth for natural languages is not possible.

On the other hand, he shows how to build such a characterization for formal languages. For a start, let us introduce a distinction between an object language and a metalanguage. Intuitively, the object language will be the language we speak about. Imagine that we are interested in discussing

[7] A fragment of a letter to Yossef Balas; see Gödel, K. (2003), p. 10.
[8] Carnap (1948), pp. vii – viii.

the properties of arithmetical sentences, with symbols for addition and multiplication. In such a discussion the language of arithmetic will play the role of the object language – the one spoken of. But we will need also a language which will be *used* (not merely mentioned) by us in such a discussion. This will be our metalanguage: the language we use to speak about the object language.

As an additional example, consider a situation when we discuss in English about the properties of German sentences. Then we have:

Es regnet heute – object language (i.e. German).

The sentence "Es regnet heute" is true – metalanguage (English).

The sentence "Es regnet heute" is composed of 3 words – metalanguage (English).

In what follows I will assume that our metalanguage contains names of all the expressions of the object language (a device like quotation could do the trick). Unlike in the above example, I will assume also that the metalanguage contains the object language itself – that every sentence of the object language belongs also to the metalanguage, but not the other way round.

On Tarski's approach the truth-bearers are sentences (cf. question (a) discussed earlier). More specifically, the predicate "true" will be treated as applying to sentences of a fixed formal object language *L*, which does not contain its own semantic expressions (it means in particular that phrases like "true sentence of *L*" do not belong to *L*). Such a restriction allows us to consider typical formal languages; note however that it excludes natural languages, like English, from the scope of investigations: clearly the expression "true sentence of English" is itself an English expression.

Right at the start Tarski introduces two conditions which the definition of truth should satisfy: any acceptable definition should be formally correct and materially adequate.

- Formally correct: it should be an explicit definition (a definition of the form "A sentence *S* is true if and only if …") not using previously undefined semantic concepts on the right side.
- Materially adequate: it should capture the meaning of our intuitive concept of truth.

The formal correctness condition is pretty standard; its fulfillment permits us to avoid typical definitional mistakes. Certainly we should not use poorly understood concepts in definitions, and since there have been doubts about the intelligibility of semantic notions, any definition of truth employing them and taking them for granted would fall short of its aim.

On the other hand, the material adequacy condition as stated above is rather vague. Fine, so we would like to capture the meaning of our ordinary notion of truth, but what does it mean to "capture" such a meaning? It is at this point that Tarski makes his famous move, changing the vague material adequacy postulate into a condition with a quite concrete mathematical content.

Material adequacy condition explained: the definition of truth (together with the axioms of the metatheory in which we formulate this definition) should imply all the so called T-sentences, i.e. all the biconditionals of the form:

(T) *S* is true if and only if *p*

with a name of a given object language sentence substituted for "*S*" and this very sentence substituted for "*p*".

Let us see some examples of concrete T-sentences:

"Snow is white" is true if and only if snow is white.

"2 + 2 = 4" is true if and only 2 + 2 = 4.

"The queen of England is wise" is true if and only if the queen of England is wise.[9]

The intuition here is that truth, in concrete applications, functions as a device of disquotation: to call a sentence true is equivalent to repeating this very sentence. This is the trait of our intuitive notion of truth which Tarski wants to preserve in his formal construction. This is also one of the basic reasons why we treat his definition as a definition of *truth*, and not of something else.

Truth definition: a general idea

Tarski defines truth from another semantic notion, called "satisfaction". Satisfaction should be conceived of as a relation between objects in the world and formulas of the object language. It should be stressed that the notion of a *formula* used here is more general than the notion of a *sentence* – in fact for typical formal languages, all sentences will be formulas, but not all formulas are sentences. Consider again the language of elementary arithmetic as an example. The set of expressions available to us in this language contains (among other things) variables (for this role we may use letters like *x*, *y*, *z* etc.) and two quantifiers, written usually as "$\forall$" and "$\exists$". The first of them is the general quantifier, read as "every" or "for all"; the second is existential quantifier, read as "for some" or "there is". The expressions below are examples of arithmetical sentences:

(a) $\exists x\,(x + x = x)$

(b) $\forall x\,(x + x = x)$.

They obtain the following natural reading: (a) there is a number *x* such that *x* added to itself gives *x* as a result; (b) every number *x* added to itself gives *x* as a result. (We see easily that the first sentence is true, with 0 being an example, and the second is clearly false.)

[9] If the examples seem trivial or non-informative, just keep in mind our assumption that the object language is contained in the metalangue. For comparison consider the T-biconditional:

"Es regnet heute" is true if and only if it is raining today,

which could be in fact quite informative for someone who does not speak German.

Both (a) and (b) are examples of arithmetical *sentences*: whenever a variable is used, it is specified whether it concerns all objects (in this case the occurrence of the variable is preceded by the general quantifier) or some objects (then it is preceded by the existential quantifier). In such a situation we say that all the occurrences of the variables are *bound* by a quantifier. On the other hand, some arithmetical expressions, although grammatical, contain free, i.e. unbound variables:

(c) $x + x = x$

We say that the expression (c) is a formula, but not a sentence. Taken by itself, it is neither true nor false: we do not know what *x* is and how it should be interpreted. It will become however true or false after the variable becomes bound by a quantifier, i.e. after we attach a quantifier "there is" or "every" at the front of the whole formula (c). In general, sentences are formulas which do not contain free variables.

As we said, in order to characterize truth, Tarski defines first a (semantic) relation of satisfaction, which holds between formulas – possibly with free variables – and sequences of objects in the world. The intuitive idea is as follows. Consider a formula with two free variables "*x* is taller than *y*". Consider the two element sequence (John, George). Now we say that:

> The formula "*x* is taller than *y*" is satisfied by the sequence (John, George) if and only if John is taller than George.

It turns out that it is possible to generalize it to cover arbitrary formulas of the formal object language. Assume for example, that at a given stage we consider a sequence σ and a formula *F* of the object language, which has the form "φ and ψ". Assume also that we have the notion of satisfaction already explained for formulas simpler than *F* (so we know in particular what it means for φ and ψ, taken separately, to be satisfied by a sequence). Then we can define:

> *F* is satisfied by σ if and only if φ is satisfied by σ and ψ is satisfied by σ.

As we stressed, this sort of treatment can be extended to arbitrary formulas of the object language, making satisfaction a fully defined notion.

Truth is defined then in terms of satisfaction. It turns out that a sentence of the object language is satisfied either by all sequences or by none. This observation depends on a technical lemma, with the following intuitive content: the formula's satisfiability by a given sequence depends on the interpretation (provided by the sequence) of the free variables in this formula. Consider the formula "*x* is a millionaire" – it will be satisfied by sequences with a millionaire as the first element, all other elements being irrelevant since there are no additional free variables to be interpreted with them. In case of a sentence (which is – we remind – a formula without free variables), the sequence as a whole becomes simply irrelevant. The sentence "George is taller than John" is satisfied by an

arbitrary sequence – say, (Rebecca, Daniel, Greta) – if and only if George is taller than John. Elements of the sequence do not matter in this case since a sentence contains no free variables for which they could provide interpretation. In effect the following definition can be adopted:

A sentence is true if and only if it is satisfied by all objects.

It turns out that a truth definition constructed in this way satisfies both formal and material adequacy condition. Truth *is* a respectable scientific concept after all!

The Liar paradox revisited

We remember the problematic sentence (L) stating its own falsity, which gave rise to the liar paradox. Now, with the Tarskian truth predicate at hand, we can ask again: is (L) true or false?

Let *True* be the Tarskian truth predicate for the object language *J*. The liar sentence (L) is viewed now as saying:

(L) is not *True*.

In order to analyze the behavior of (L), we make the following observations:

1. (L) does not belong to the object language *J*.
2. So (L) is not *True*.
3. Therefore (L).

It does not follow however that (L) is *True*.

Comments. Observe first, that (L) contains the Tarskian truth predicate, therefore it does not belong to the object language *J*, which by assumption does not contain semantic expression. This observation is the content of 1 above. We know in addition that only sentences of the object language *J* fall under the predicate *True*. Therefore (L) indeed is not *True* (see 2 above). Since that's what (L) states, in step 3 we conclude that (L).

At this moment we are tempted to think that in such a case (L) is intuitively true – it turned out after all that (L) states something that really obtains! This temptation should be resisted. In Tarski's framework we are not entitled to the conclusion that (L) is *True*. In fact (L), although asserted by us in step 3, is not *True* – as we said, only sentences of the object language *J* can be *True* and (L) does not belong to *J*. In this way the contradiction is avoided, but the intuition about the truth of (L) admittedly remains with us. This intuition can be expressed however in Tarskian framework: we can introduce a new predicate $True_1$ for the new object language, namely for the old language *J* extended with the *True* predicate, and repeat the whole construction. In other words, we climb to the next level, treating our previous metalanguage as a new object language in this new stage. Eventually the

sentence (L) will come out as $True_1$, i.e. true in the sense of the truth predicate of the next level in the Tarskian hierarchy.

(Un)definability of truth

Alfred Tarski showed how to define truth but he proved also a famous theorem of undefinability of truth (which in the literature is called quite often just "Tarski's theorem"). In broad, informal terms, the theorem states that truth is undefinable. We saw however that Tarski provided a definition of truth; so how can truth be undefinable if it has been defined by Tarski?

The answer lies in a careful formulation of Tarski's theorem. In a narrow version, the theorem states that arithmetical truth cannot be defined in arithmetic. More generally, it states that truth for the object language cannot be defined by means available to us in theories formulated in the object language itself. In other words: if we want to define truth for the object language, we need richer means, going beyond the object language itself. Indeed, Tarski showed how to construct such a definition in a richer metatheory.

We arrive finally at the explanation of the title of this paper: truth, undefinable in the object language, can be defined in a richer metatheory. Tarski showed how to do this, initiating a whole new area of research, called nowadays a "model theory".

Acknowledgements. The author was supported by a grant from the National Science Centre in Cracow (NCN), decision number DEC-2011/01/B/HS1/03910.

BIBLIOGRAPHY

[1]. Banach, S., and Tarski, A. (1924), "Sur la décomposition des ensembles de points en parties respectivement congruentes", *Fundamenta Mathematicae* 6, pp. 244-277.

[2]. Carnap, R. (1948), *Introduction to Semantics*, Cambridge, Massachusetts, Harvard University Press.

[3]. Feferman, A., and Feferman, S. (2004), *Alfred Tarski. Life and Logic*, Cambridge University Press, Cambridge.

[4]. Gödel, K. (2003), *Collected works*, vol. IV, Oxford University Press, Oxford.

[5]. Tarski, A. (1923), "O wyrazie pierwotnym logistyki", Doctoral dissertation, published in Przegląd Filozoficzny, vol.26, pp. 68–89. An English translation appears in (Tarski 1983), pp. 1–23, under the title "On the primitive term of logistic".

[6]. Tarski, A. (1933), “Pojęcie prawdy w językach nauk dedukcyjnych”, Towarzystwo Naukowe Warszawskie. Translated by J.H. Woodger as “The concept of truth in formalized languages”, in Tarski (1983), pp. 152-278.

[7]. Tarski, A. (1944), “The semantic conception of truth and the foundations of semantics”, *Philosophy and Phenomenological Research* 4, pp. 341–376.

[8]. Tarski, A. (1983), *Logic, semantics, metamathematics. Papers from 1923 to 1938*, (translations by J. H. Woodger), Hackett Publishing Company, Indianapolis, Indiana.

[9]. Woleński, J. (1989), *Logic and Philosophy in the Lvov-Warsaw School*, Kluwer Academic Publisher, Dordrecht.