When Snow Is White
Tarski's Theory of Truth

Y. Duppen
yduppen@xs4all.nl

April 7, 2002

1 Introduction

Since the dawn of philosophy philosophers have argued about the concept of truth. Questions like: "What does it mean for something to be true?" and "Is there an absolute truth?" resulted in all kinds of theories. As befits philosophical theories, they heartily disagree with each other.

This paper is concerned with the Semantic Theory of Truth, which is devised around 1930 by Alfred Tarski. His theory was different in this aspect that it did not explain truth directly; instead, it reduced the concept of truth to other semantical notions. According to Tarski this reduction would be neutral, and therefore applicable to all truth theories. Chapter 2 will give an overview of this theory with its implications. A more formal definition will be given in chapter 3.

Of course, Tarski's theory met some criticism. Chapter 4 will show that the theory is not as neutral as Tarski would like us to believe, but has instead been influenced by physicalism.

Among the critics is Henry Field who, instead of merely criticizing Tarski's theory, tried to *improve* the theory. Chapter 5 deals with Field's criticism: Field's theory "Tarski*" will be explained, together with a short comparison between Tarski's theory and "Tarski*".

2 Tarski's Semantic Theory of Truth

Before going into Tarski's theory of truth, it is important to realize what exactly it tries to accomplish. In accordance with most theories of truth, it does not aim to explain *truth* as a stand-alone concept. Instead, it tries to define "what it is for a proposition to be true"[1].

One of the major weaknesses of other theories, is that they do not hold when related to paradoxes. Even the ancient Greeks faced this problem, wondering for instance how to cope with Eubulides' Paradox[6]:

> This statement is false

In order to find a theory that could deal with paradoxes, Tarski reasoned that it was necessary to find the cause of the paradox. Like Russel[6], he found that the problem lies in the fact that the statement reasons about itself. The sentence is part of a semantically closed language, a language that can refer to itself and reason about itself.

So, in order to reason about the truth of a sentence, it is necessary to distinguish the language *about* which we reason from the language *in* which we reason. In other words, we should make a distinction between the object language and the metalanguage. The object language should not be semantically closed, while the metalanguage should have the ability to refer to sentences in the object language. This distinction does not solve the problem of paradoxes as such. It merely prevents a definition of truth to be applied to a paradox.

For Tarski the required distinction between object language and metalanguage had an interesting consequence: he deemed it impossible to give a truth definition for natural languages. After all, natural languages are all semantically closed. Furthermore, they are usually ambiguous and vague. And finally, natural languages are continually changing, making it impossible to give a fixed definition of truth. As Tarski said himself:

> The problem of the definition of truth obtains a precise meaning and can be resolved in a rigorous way only for languages whose structure has been exactly specified.[5]

Apart from tackling the problem of the paradox, Tarski also wanted his theory to be formally correct and materially adequate[5]. The latter condition goes back to Aristotle:

> *...το μϵν γαρ λϵγϵιν το όν μη ϵ́ιναι ή το μη όν ϵ́ιναι ψϵυδος,*
> *το δϵ το όν ϵ́ιναι και το μη όν μη ϵ́ιναι άληθϵς....*

> 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, or of what is not that it is not, is true.[7]

Tarski formulated the requirement of material adequacy in his famous convention (T):

> (T) “p” is true in L, iff p.

In order to understand this, it is necessary to see the difference between “p” and p. “p” is the sentence stating a certain proposition in a certain object language L. On the other hand, p is the translation of that sentence into the metalanguage. It is not an utterance but a possible state of affairs.

For example, if we take Dutch as our object language and English as our metalanguage[1] , an instance of convention (T) could be:

> “Sneeuw is wit” is true in Dutch, iff snow is white.

This convention has received a lot of criticism. Chapter 4 deals with some of it. There is however one important point that should be made here: convention (T) is not a definition. It does not determine the meaning of *true*, but the extension — all cases to which *true* should apply. In other words: if we find a definition of truth which does not comply with (T), the definition is wrong.

Having identified the preconditions of a definition of truth, Tarski could now proceed with the actual definition of truth. Instead of an isolated definition, Tarski chose to define it in terms of an ‘easier’ semantical concept: satisfaction.

To understand the notion of satisfaction, some background in formal languages is necessary. All formal languages can be described by an (inductive) grammar. This is done firstly by giving the base objects, followed by combining rules. The grammar in table 1 is an example of such a grammar. In this grammar, the rules 1, 2 and 3 define the base objects. The rule 4 inductively defines the combining rules.

With this grammar, we can generate sentences like “snow is white”, “water is blue” or “snow is white and corn is white” by starting with the $\langle sentence \rangle$ object. However, we can also generate partial sentences, or sentential functions, by stopping halfway. An example of such a sentential is “snow is $\langle adjective \rangle$”.

[1] Strictly speaking we can not do this since the semantic theory of truth only applies to formal languages. I believe, however, that examples in natural languages are easier to comprehend. Therefore most examples will be in English.

$$\begin{aligned}
\langle \mathit{adjective} \rangle &\rightarrow \texttt{white} \mid \texttt{blue} \mid \texttt{yellow} && (1)\\
\langle \mathit{noun} \rangle &\rightarrow \texttt{snow} \mid \texttt{water} \mid \texttt{corn} && (2)\\
\langle \mathit{verb} \rangle &\rightarrow \texttt{is} && (3)\\
\langle \mathit{sentence} \rangle &\rightarrow \langle \mathit{sentence} \rangle \ \texttt{and} \ \langle \mathit{sentence} \rangle && \\
& \quad \mid \langle \mathit{noun} \rangle \langle \mathit{verb} \rangle \langle \mathit{adjective} \rangle && (4)
\end{aligned}$$

Table 1: A simple grammar

Using this rigid structure, Tarski could define satisfaction as relation between objects and sententials, after which he could give a definition of truth in terms of satisfaction. Intuitively, satisfaction should relate those objects and those sententials that together form a true sentence. For instance, "white" should satisfy the sentential "snow is $\langle \mathit{adjective} \rangle$", because snow is white. However, the definition of satisfaction should not include words like 'true'.

Tarski solved this by following the structure of the grammar. Firstly all base sententials are enumerated. For those sententials satisfaction is defined directly, after which satisfaction is defined for the combining rules. For instance, the first part of rule 4 in the example grammar gives rise to the following definition of satisfaction: if "O_1" satisfies "S_1" and "O_2" satisfies "S_2", then $\langle$"O_1","O_2"$\rangle$ satisfies "S_1 and S_2".

There is now only one step left to get to Tarski's definition of truth; for this we have to look at sentential functions without 'open places'. These are the sentences. Because there are no open places, sentences can either be satisfied by all objects or by no objects. So, if there is a sensible definition of satisfaction for the example grammar, all object satisfy the sentence "snow is white", while there is no object that will satisfy the sentence "snow is blue". These observations resulted in Tarski's definition of truth:

> "A sentence is true iff it is satisfied by all objects and false otherwise."[5]

3 A More Formal Approach

Tarski not only wanted a materially adequate definition of truth; he also aimed for a formally correct one. This chapter will prove that he is right by giving a formal account of Tarski's theory of truth.

In order to solve the problem in a materially adequate way, Tarski saw two possible solutions. The first approach is to define the semantic concepts by axiom. This would cause the semantics to be independent of the object language. However, Tarski saw several problems with this solution, among which the choice of the axioms. He held the opinion that this choice is always arbitrary, and therefore invalid. Besides that, it is difficult to choose a set of axioms such that they remain consistent, sound and complete. And finally, a theory of truth based on arbitrary axioms does not fit into physicalism[4][2].

Because of these problems, Tarski decided to use the other solution: reducing semantic axioms to logical concepts, concepts from the object language and morphological concepts.

Tarski argued that it is impossible to talk about the truth-value of an arbitrary sentence. Truth only makes sense with respect to a given (formal) language, so we can only talk about the truth of a sentence *in a specific language*. This language, called the object language L, has to be different from the metalanguage M in which the truth-value is determined. Furthermore, L has to be a true subset of M, enabling M to refer to sentences from L. In [4], Tarski gives the following algorithm:

1. Describe L by enumerating the primitive terms and giving an inductive definition
2. Find the set of sentences S that can be constructed in L, followed by a set $A \subseteq S$ of axioms
3. Formulate the rules of inference
4. Construct M in such a way that it contains:
    (a) The object language L
    (b) Expressions to deal with the morphology of L
    (c) Logical expressions

This methodology makes it easy to define satisfaction. Mathematically speaking, satisfaction is a relation between sententials from L and objects, expressed in M. More precise, each sentential in L can be seen as an n-ary function $Fx_1 \dots x_n$. The satisfaction relation is then a relation between those functions and infinite sequences of objects, projected to sequences of n elements. Now *true* can be defined as 'satisfied by all sequences', while *false* can be defined as 'satisfied by no sequences'.

The inductive definition of L makes it possible to define this relation. By enumerating the primitive terms (the axioms), one can define the satisfaction relation for each of these terms. And since M has the ability to refer to the morphology of L, the satisfaction relation can also be defined for the rules of induction.

[2] More in this statement in section 4

4 Criticism of the Theory of Truth

Tarski claimed that his theory was a neutral theory: because it is only used in a formal context, it does not use epistemological or metaphysical terms. Therefore, he argued, his theory could be used in connection with any other theory.

However, when having a closer look at the theory, it turns out that it does possess some physicalist notions. The most obvious is the satisfaction of sententials: sententials can be satisfied by objects, things that exist in our world. Satisfaction by objects excludes a large class of theories, namely those using satisfaction by *name*. The requirement that these objects are located in our world excludes Kripke-style possible world models.

Related to this is the use of the existential quantifier. Tarski proposes an objectual reading of this quantifier, again ignoring the possibility of names (substitutional quantification).

Another touch of physicalism can be found in convention (T). The condition for material adequacy only works for bivalent truth systems, i.e. truth systems that only use the values *true* and *false*.

The final proof of Tarski's preference for physicalism is his own statement in [4], when he dismissed a certain approach to a definition of truth:

> . . . it would be different to bring this method into harmony with the postulates of unity of science and of physicalism. . . .

However while Tarski's work is not neutral, as Tarski himself believed, it is still important. As Field says:

> Tarski succeeded in reducing the notion of truth to certain other semantics [3]

In order to understand this quote, we have to look at it from a historical perspective. As can be seen from the vast number of theories of truth, defining truth has always been problematic. In the 1930's, certain physicalists who failed to give a satisfying definition of semantic notions, decided to throw away those notions. Tarski proved them wrong, by showing that (for certain languages) it is possible to explain these notions non-semantically.

Apart from the claim of physicalism, some people have claimed that convention (T) is a sign of realism. But Tarski pointed out that (T) does not necessarily refer to the real world; for instance, '"snow is white" is true iff snow is white' only says that if we reject 'snow is white' for one reason or another, we should also reject '"snow is white" is true'.

The final point of criticism that will be dealt with here is the contents of Tarski's 'truth'. Some people pointed out that his theory only covers specific areas of truth. Tarski admitted this; his theory is not about truth in general, but about what he calls *fruth*, truth in a formal context.

5 Enhancements of Tarski's Theory

While the criticism in the previous section is arguable$^?$, Hartry Field noticed a far more interesting problem in Tarski's theory. Recall from section 2 that a truth definition starts by finding the base sentences and the combining rules, to which truth values are assigned. Tarski suggested this should be done by enumerating them. Field notes that such a definition by enumeration has several shortcomings [3].

First of all, this means that the object language should not contain ambiguous names. For if it does, only one meaning will show up in the truth definition; sentences that use the other meaning will now be assigned a 'wrong' truth-value.

Secondly, it only works for languages in which nothing is denoted that cannot be denoted in the metalanguage. Some people would not consider this a shortcoming, since Tarski's 'recipe' for constructing the metalanguage ensures that the metalanguage can denote everything the object language can denote (see section 3). Field was however interested in extending this theory as much as possible, and I agree that it is a valid point.

Finally, and most importantly, definitions by enumeration can not apply to evolving languages. Every time the language changes, the definition has to be changed. Of course, for a fixed formal language this is no problem, but if you want to extend the theory to natural languages (as Field did), it becomes a serious obstacle.

Field tries to solve this problem by creating the theory Tarski* (table 3). In order to compare his Tarski* theory with the original theory, he formalizes Tarski's theory in the same manner (table 2). He then proceeds by convincing the reader that Tarski* is a better theory than Tarski. [3]

When we compare both definitions, we see that they are almost identical. The only real difference lies in the denotation of constants. All other differences he points out are merely special cases of this. Of course this is partially due to Field's rewriting of Tarski's definition. As said before, Tarski did not explicitly use the concept of denotation. Instead he defined constants, predicates and functions by enumeration. The reason for this is that he desired

[3] Actually, Field goes one step further. He not only tries to convince the reader that Tarski* is better than the original, but he also tries to convince us that Tarski* is the theory Tarski actually had in mind.

denotation$_s$:

1. 'x_k' denotes$_s$ s_k
2. 'c_k' denotes$_s$ $\overline{c}_k$.
3. $[f_k(e)]$ denotes$_s$ an object a iff
   (a) there is an object b that e denotes$_s$
   (b) a is $\overline{f}_k(b)$

true$_s$

1. $[p_k(e)]$ is true$_s$ iff
   (a) there is an object a that e denotes$_s$
   (b) $\overline{p}_k(a)$.
2. $[\neg e]$ is true$_s$ iff e is not true$_s$
3. $[e_1 \wedge e_2]$ is true$_s$ iff e_1 is true$_s$ and e_2 is true$_s$
4. $[\forall x_k(e)]$ is true$_s$ iff for each sequences $s*$ that differs from s at the kth place at most, e is true$_s$

A sentence is true iff it is true$_s$ for some (or all) sequences of objects s.

Table 2: Tarski's definition of truth[3]

denotation$_s$:

1. as in table 2
2. 'c_k' denotes$_s$ what it denotes
3. $[f_k(e)]$ denotes$_s$ an object a iff
   (a) as in table 2
   (b) 'f_k' is fulfilled by $\langle a, b\rangle$

true$_s$

1. $[p_k(e)]$ is true$_s$ iff
   (a) as in table 2
   (b) 'p_k' applies to a

2. $-$ 4. as in table 2

A sentence is true iff it is true$_s$ for some (or all) sequences of objects s.

Table 3: The Tarski* definition of truth[3]

“semantic terms (referring to the object language) to be introduced into the metalanguage only by definition.”[3], to avoid circular definitions.

Because Field does not consider definition by enumeration adequate, he has no choice but to introduce another term. He argues convincingly that the concept truth is not necessary to define denotation[4]. Basically this is all the difference there is: denotation is defined without a reference to truth, while truth is defined without referencing a *specific* denotation. This enables us to change the denotation without having to change or extend the definition of truth. This makes the Tarski* theory even more neutral than the original.

6 Final Notes

In short the Semantic Theory of Truth gives a satisfactory account of defining truth for formal sentences by making use of their inductive structure. Its validity has been proven by explaining some of the formal aspects of the theory. But while Tarski claimed that his theory was completely neutral, chapter 4 shows that for a neutral theory it contains too many physicalist notions. Finally Fields improvement on Tarski's theory, the removal of definition by enumeration, has been discussed.

I have tried to explain the theory as extensively as possible, but several points should be noted.

- Davidson has done a lot of work in extending Tarski's theory. A good reference for this is [2].
- Field's ideas on denotation and satisfaction go much deeper. See [3] for more details.
- There has been a lot more criticism on Tarski than just his preference for physicalism. A brief account of this criticism, together with Tarski's objections can be found in [4]

[4] After all, Tarski's definition by enumeration is an example of a definition that does not use truth

References

[1] A.C. Grayling: *An Introduction to Philosophical Logic*, 3rd edition, 1997, pp. 147-160

[2] D. Davidson: Truth and Meaning, in *???*, pp. 17-36

[3] H. Field: Tarski's Theory of Truth, in *Journal of Philosophy, lxix, 13*, 1972, pp.347-375

[4] A. Tarski: The Establishment of Scientific Semantics, in *???*, pp. 401-408

[5] A. Tarski: The Semantic Conception of Truth, in *Philosophy and Phenomenological Research*, pp. 341-375

[6] E. W. Weisstein: *World of Mathematics*,
`http://mathworld.wolfram.com`

[7] Aristotle: *Metaphysics* 1011 b 26,
The Perseus Project: `http://www.perseus.tufts.edu`