A brief exploration of Tarski’s theory of truth and its relation to the liar paradox.

“This sentence is not true.”

Most literate and logically-oriented people would agree that a meaningful sentence is always either true or false, but never both, and very few speakers of English would deny that the sentence above is meaningful. But assuming a commonsense notion of truth and a logical principle of bivalence,1 we can easily derive a contradiction:

- the sentence asserts that it isn’t true – in other words, that it is not an accurate picture of reality;
- either this is the case or it is not;
- if this is the case and the sentence is not true, then since what the sentence asserts happens to be the case, the sentence is an accurate picture of reality and therefore true; thus if the sentence is not true, then it is true;
- if this isn’t the case and the sentence is not not true, then since what the sentence asserts happens not to be the case, the sentence is not an accurate picture of reality and therefore not true; thus if the sentence is true, then it is not true.

We are confronted with two tasks: (1) to somehow account for the truth value of the liar sentence2 without violating bivalence, and (2), to define truth more exactly.

The circularities inherently involved in the pursuit of a definition of truth have pushed many theorists from the realm of natural language into the realm of the formal. Here, definitions and theories can be comfortably developed and tested on fragments of natural language or small artificial languages built or delineated specifically in order to avoid ambiguities of meaning and truth. Alfred Tarski is a modern exemplar: the formal consistency of his theory has provided an historically and conceptually important foothold for subsequent theories of truth.

Tarski, largely held to be among the greatest logicians of all time, was born in Warsaw in 1901, and it was there in the years before his 1939 emigration to the United States (where he eventually became a fixture at Berkeley), while supporting himself primarily by teaching high school mathematics, that he published many of his early seminal papers in logic and mathematics. Among them was the 1933 “Pojęcie prawdy w językach nauk dedukcyjnych” (“The concept of truth in the language of the deductive sciences”), which singularly established the basis for logically sound semantics by setting up a theoretical structure in which truth (or, Tarski’s version of truth) can be consistently comprehended.3

---

1 Principle of bivalence: for any proposition P, (1) either P is true or P is false, (2) either P is true or not-P is true, and (3) it is not the case that both P is true and not-P is true. I also assume that, for any proposition P, P is true iff what P “expresses” “is the case” or “is an accurate picture of reality.”
2 The nickname “liar” is due to the original, somewhat weaker, formulation of the paradox, which is credited to Epimenides of Crete (ca. 600 BCE) and states that “All Cretans are liars.”
3 It is not “Pojęcie prawdy w językach nauk dedukcyjnych” that I have consulted for this paper but rather “The Semantic Conception of Truth,” Tarski’s briefer 1944 re-exposition in which he confines himself to a
Tarski gives two criteria for the success of a sufficient definition: (1) that the definition be *materially adequate*; in other words, that the definition ascribe truth to every true entity and no fewer; and (2) that the definition be *formally correct*; in other words, that the theoretical framework in which the definition is to be embedded establish the formal structure of the language in which claims are to be made about truth and that the definition be consistent with respect to said structure. He also notes that any time we talk about the truth of a sentence, we are implicitly talking about its *truth-in-a-particular-language*: for, “the same expression that is a true sentence in one language can be false or meaningless in another.” (1944 p 342)

In order to know whether criterion (1) holds, it must first be known of which objects it is allowed to predicate truth and falsity. Having considered the potential truth-bearing nature of propositions, thoughts, and beliefs, Tarski restricts the domain of his theory to sentences, but not without suggesting that it might later be extended to include such other types of objects.

In considering which intuitions about truth we should expect a good definition to satisfy, Tarski cites Aristotle’s formulation of correspondence:

<table>
<thead>
<tr>
<th>STATEMENT</th>
<th>“x is”</th>
<th>“x isn’t”</th>
</tr>
</thead>
<tbody>
<tr>
<td>x is</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>x isn’t</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Under this system a sentence is true if and only if it “corresponds to reality” or “designates an existing state of affairs.” Tarski observes that this notion is not a sufficiently precise or clear one on the basis of which to formulate a careful definition and suggests that we consider how we might specify more precisely the conditions under which, for example, the sentence “Snow is white.” is true. The most obvious formulation is “Snow is white.” is true iff snow is white.

Here the use of “snow is white” on the right is one way to describe what state of affairs must hold in order for the sentence to be true, and in fact the most precise one, which happens to appear identical to the sentence itself. The mention of the sentence in quotations on the left is one way, among many, to name the sentence. If we abstractly consider an arbitrary sentence, replace it with $p$, and an arbitrary name for it, $X$, then we have the following general schema:

$$X \text{ is true iff } p.$$
Our definition will be materially adequate if it entails a statement of this form for every possible value of \( p \).

Turning now to the criterion of formal correctness, notice that the problem of the definition of truth is only meaningful for exactly specified languages – languages for which we have an unambiguous characterization of the class of meaningful sentences. Since by most accounts it appears as though nothing of the sort is possible for a natural language such as English, it would be idle to attempt a formally correct definition with respect to English. Instead, Tarski suggests, we might try to approximate English as closely as possible using some exactly specified language.

Tarski next presents an ominous antinomy the implications of which he believes must be considered in the development or delineation of the exactly specified language for which truth is to be defined. This antinomy is, of course, the paradox of the sentence that declares itself false. Since plugging the sentence as it was formulated above directly into the truth schema would necessarily shift references, let us follow Tarski in reformulating it:

“The sentence printed in this paper on page 3, line 16, is not true.”

---

4 Granted that any language of interest will probably contain an infinite number of sentences, our definition now appears to have the structure of an infinite disjunction.

5 Assuming a language to include an infinite number of sentences, a useful semantic theory must operate on a principle of compositionality: for the theory to rest on a finite axiomatic base, the meanings (and therefore truth values) of complex sentences must be systematically derived from the meanings of their parts. A successful semantic theory accomplishes this by giving a finite set of axioms attributing meaning to the smallest atomic units of the language and a finite set of recursive rules by which the meanings of larger expressions can be derived from the meanings of their smaller components. Different theories have different thresholds, as it were, for meaningfulness, some ascribing meaningfulness only to sentences and reducing the meaning of words and non-sentential phrases to merely the systematic effects they have on the meanings of sentences, others giving meanings of different types, but meanings nonetheless, to all syntactic “units” (Davidson’s theory is an example of the former, Frege’s, the latter).

6 Without this characterization we would be in a similar position to, e.g., a person instructed to collect all the white buttercups in a certain area, but having no idea what constitutes a buttercup.

7 For, in English, “We do not know precisely which expressions are sentences, and we know even to a smaller degree which sentences are to be taken as assertible.” (Tarski 1944 p 349) For example, the sentence “I might could go to the store.” would be judged assertible by some speakers of American English but not others. Such variation in assertibility essentially does not arise for fully formalized languages.

8 And this is precisely what Montague tried to do.

9 Although a definition of truth formulated along Tarski’s lines will not succeed in capturing truth consistently and completely with respect to a natural language, it can certainly succeed with respect to the class of formalized languages – languages in specifying the structure of which we refer exclusively to the form of the expressions involved. Tarski is eager to point out that “entire branches of science” can be formulated within formalized languages: mathematics, logic, and theoretical physics, for example (1944 p 347).

10 “This sentence is not true.” is true iff this sentence is not true.
Plugging *this* sentence into the schema yields

“The sentence printed in this paper on page 3, line 16, is not true.” is true iff the sentence printed in this paper on page 3, line 16, is not true.

Since the sentence printed in this paper on page 3, line 16, is “*The sentence printed in this paper on page 3, line 16, is not true.*”, we have a direct contradiction, to further perspicuate which we can replace the sentence with a symbol $s$:

$s$ is true iff $s$ is not true.

Now the question becomes which of our premises is responsible for generating the contradiction. Tarski identifies two as being necessary to generate it: (1) that the language in which the antinomy is formulated is semantically closed, *i.e.* that it contains in addition to its expressions also the names of these expressions and semantic terms such as “true” referring to sentences of the language; and (2) that in the language, the “ordinary laws of logic” hold (1944 p 348). Since rejecting premise (2) would have relatively unmanageable consequences, Tarski opts instead to reject premise (1) and posits that a consistent definition of truth can only be obtained for a language that is not semantically closed. Therefore, in generating an exactly specified approximation of English, we must be careful not to allow the generants of semantic closure.

Now in order to talk about the truth of the sentences of some language, we have no choice but to use a second language. Here are introduced the relative notions of *object language* and *meta-language*, the former being the language the truth of whose sentences we are concerned to define, and the latter being the language in which we formulate the definition. The implicit boundedness of truth to a particular language here becomes useful for us to make explicit: say we have a non-semantically-closed, exactly specified approximation of English – call it English$_1$ – and a non-semantically-closed, exactly specified approximation of English containing in addition to all the expressions of English$_1$, names for them, and the semantic predicate true-in-English$_1$ – call this language English$_2$. With English$_1$ as object language, we can now use meta-language English$_2$ to formulate truth statements for English$_1$. Since English$_2$ is a *superset* of English$_1$, all the sentences of English$_1$ that appear as statements on the right sides of schema-form biconditionals will have their proper translations in English$_2$:

“Snow is white.” is true-in-English$_1$ iff snow is white.

“Grass is green.” is true-in-English$_1$ iff grass is green.

etc.

We now have a consistent definition of truth, and an attempted formulation of the liar sentence in this system becomes nonsensical:

“This sentence is not true-in-English$_1$.”
Since it is necessarily not a sentence of English\textsuperscript{1}, it is nonsensical in the same way “Snow est blanche.” is nonsensical. Similarly, “This sentence is not true-in-English\textsubscript{2}.” has no truth value in English\textsubscript{2} because it is necessarily a sentence of English\textsubscript{3} (or possibly English\textsubscript{4} or English\textsubscript{5}, and so on), and so on. The hierarchy renders the liar sentence neither true nor false but rather completely nonsensical.

To deny the liar sentence any sense at all is only one of several ways to evade the paradox; one could conceivably challenge any one or more of the other assumptions that, together with semantic closedness, generate the contradiction. Richard Kirkham (1992) suggests that the premises of any derivation of the liar paradox may essentially be reduced to the following:

(1) The liar sentence says that it is false and says nothing else.
(2) If a sentence says something is the case, and it is the case, then the sentence is true.
(3) If a sentence is true and says something is the case, then it is the case.
(4) Every sentence is either true or false.\textsuperscript{11}

Tarski’s assumptions (1) and (2) correspond roughly to Kirkham’s (1) and (4), respectively.\textsuperscript{12} Using Kirkham’s template as a frame, it appears that Tarski chooses simply to abandon Kirkham’s premise (1), claiming that the liar sentence in fact says not only that it is false but that it is false-in-the-object-language.

Any one of Kirkham’s premises, though, is eligible to be challenged and modified in the construction of a new theory that might just as successfully circumvent the paradox. Saul Kripke, Graham Priest, and Bradley Dowden have all challenged (4), for example. Kripke suggests that the liar sentence is one of a class of statements that, in certain circumstances, simply have no truth value at all, and modifies his logic to account for such a truth value “gap.” Priest and Dowden both deny the logical principle of explosion, which allows anything to follow from a contradiction, as a valid rule of inference, and suggest that the liar sentence is both true and false simultaneously. A.N. Prior challenges premise (1), and suggests that every sentence implicitly predicates truth of itself, thus rendering the liar sentence a contradiction, and thus false, from the moment it is uttered.

\textsuperscript{11} The rest of the derivation goes like this:
(5) The liar sentence is true. (Conditional premise.)
(6) Ergo, the liar sentence is false. (By 1, 2, 5.)
(7) Ergo, the liar sentence is true and false. (By 5, 6.)
(8) Ergo, if the liar sentence is true, then it is true and false. (By conditional proof.)
(9) The liar sentence is false. (Conditional premise.)
(10) Ergo, the liar sentence is true. (By 1, 3, 9.)
(11) Ergo, the liar sentence is true and false. (By 9, 10.)
(12) Ergo, if the liar sentence is false, then it is true and false. (By conditional proof.)
(13) Ergo, the liar sentence is true and false. (By 4, 8, 12.)

\textsuperscript{12} Tarski’s “ordinary laws of logic” of course include bivalence as expressed in Kirkham’s premise (4), and while Tarski’s (1) is certainly more general than Kirkham’s (1), the semantic closedness of a language is both necessary and sufficient for generating in it a sentence that predicates falsity of itself.
It is perhaps also worth noting that the proof of the consistency of Tarski’s theory rests upon a notion developed by Bertrand Russell in his logical theory of types (1908). The theory was built to avoid a paradox similar to the liar now known as Russell’s paradox. To generate the paradox within naïve set theory, we assume (1) that every object, given some set, is exclusively either a member of it or not, and (2) that we can define a set by stating any condition and that if a given object satisfies that condition, it is a member of the set, and if it doesn’t, it is not (these roughly correspond to Kirkham’s (1), and (2)&(3), respectively). Let us now define a set \( R \) by the condition of non-self-membership:

\[
R = \{ A \mid A \not\in A \}
\]

Now the problem arises when we ask the question whether \( R \) is a member of itself: if it is, then it meets the criterion for membership and must not be a member of itself; if it isn’t, then it must not meet the criterion for membership and must not not be a member of itself. If \( R \) is a member of itself, then it is not a member of itself, and if \( R \) is not a member of itself, then it is a member of itself.

Russell’s solution is to classify sentences in a stratified way: sentences about objects are first-order, sentences about sets of objects (predicates, in other words) are second-order, sentences about sets of sets of objects (predicates of predicates) are third-order, and so on. Where Tarski blocked closure by stipulating that a given language should not contain semantic predicates which can apply to its own sentences, and that any sentence which makes a semantic predication should only do so of a sentence of a language lower than itself on the hierarchy, Russell blocks closure by stipulating that a sentence should be necessarily of one order higher than the highest-ordered entity referred to by it. Both theories structurally prevent the defining of any entity in terms of itself.

While self-reference can be identified as the structural cause of the liar paradox, most cases of self-reference in language are completely unproblematic and nonparadoxical with respect to truth. For example, “This sentence is English.” is unparadoxically true, likewise is “This sentence is five words long.” false and “This sentence is not in Spanish.” true. A theory that blocks self-reference not only blocks the paradoxes, but also numerous such unproblematic constructions.
Works cited in the text:

Works consulted while writing the text but not cited in the text: