What is Tarski’s Theory of Truth?  

Gila Sher

1. Interpretation as generalization

In “On the Concept of Truth in Formalized Languages”, Tarski (1933) describes his project as follows:

For an extensive group of formalized languages it is possible to give a method by which a correct definition of truth can be constructed for each of them. The general abstract description of this method and of the languages to which it is applicable would be troublesome and not at all perspicuous. I prefer therefore to introduce the reader to this method in another way. I shall construct a definition of this kind in connection with a particular concrete language and show some of its most important consequences. The indications which I shall then give in §4 of this article will, I hope, be sufficient to show how the method illustrated by this example can be applied to other languages of similar logical construction. (Pp. 167–168)

Tarski conceived of his theory as a general method for defining truth for a broad, if well defined, range of languages, but he chose to expound it through a single, simple example. This example, however, does not uniquely determine his general method, and the question arises as to how to generalize Tarski’s example. Tarski clarified one aspect of this question, namely, how to extend his example to languages with indefinitely high order of variables, but many other fundamental issues were not addressed either in his original (1933) paper, or, indeed, in his later (informal) papers (1944 and 1969). The fact that Tarski did not address these questions is, of course, indicative of his attitude: Tarski was either unaware of these questions, or uninterested, or believed the answers were obvious and no further explanation was required. Today, however, the philosophical discussion has veered away from the technical matters that occupied Tarski in the 30’s (partly, no doubt, due to his own thorough and successful treatment of these matters), and differences in attitude towards Tarski’s theory are often grounded in differences in answers to the open questions. Even general attitudes towards the theory of truth (e.g., towards the possibility of a substantive, non-deflationist theory of truth) can be traced to implicit generalizations of Tarski’s example.

In this paper I will study Tarski’s theory through a few of its open questions and some of its generalizations. I will concentrate on the “reductionist approach” to Tarski’s theory, exemplified by two generalizations due to Field. My critical investigation of these generalizations will not be directed at their exegetical virtues; rather, I will be interested in their viability as philosophical theories and in some of the challenges they face. I will begin with a brief introduction to the original goals of Tarski’s theory.¹

2. Aims of theory

We can distinguish three aims of Tarski’s theory: a philosophical aim, a methodological aim and a logical aim.

1. The Philosophical Aim.² Tarski described his goal in constructing a theory of truth as philosophical in nature. The goal is to construct a materially accurate and formally consistent definition of the classical notion of truth:

   The present article is almost wholly devoted to a single problem – the definition of truth. Its task is to construct . . . a materially adequate and formally correct definition of the term ‘true sentence’. This problem . . . belongs to the classical questions of philosophy . . . . [Ibid., p. 152. See also pp. 266–267]

By the ‘classical question’ of truth Tarski means the question of how to define the “classical”, correspondence notion of truth:

   [T]hroughout this work I shall be concerned exclusively with grasping the intentions which are contained in the so-called classical conception truth (‘true – corresponding with reality’) . . . . [Ibid., p. 153]
The task of constructing an adequate theory of the classical notion of truth is, however, fraught with difficulties:

This problem . . . raises considerable difficulties. For although the meaning of the term ‘true sentence’ in colloquial language seems to be quite clear and intelligible, all attempts to define this meaning more precisely have hitherto been fruitless, and many investigations in which this term has been used and which started with apparently evident premises have often led to paradoxes and antinomies . . . . (Ibid., p. 152)

Tarski divides the philosophical task into two sub-tasks: (i) the material task of capturing the exact content of the correspondence notion, and (ii) the formal task of complying with the most rigorous standards of logical consistency and correct definition. But Tarski’s treatment of these two tasks is not equal. In executing the formal task Tarski offers a substantive, in-depth analysis of the semantic paradoxes and a substantive proposal for preventing their occurrence, but in carrying out the material task Tarski offers no deep analysis of the correspondence notion of truth or the philosophical problems it gives rise to. The correspondence notion is treated either as well understood:

the meaning of the term ‘true sentence’ in colloquial language seems to be quite clear and intelligible (ibid.),

or as a notion whose analysis is to be given elsewhere.

A thorough analysis of the meaning current in everyday life of the term ‘true’ is not intended here. Every reader possesses in greater or less degree an intuitive knowledge of the concept of truth and he can find detailed discussions on it in works on the theory of knowledge. (Ibid., p. 153)

And the task is conceived as defining the bare skeleton of the philosophical notion of truth in a precise and formally correct manner, not as providing a new, deeper analysis of the material content of this notion. In this way the material task itself is construed as a formal task.

This situation creates special difficulties for a philosophical study of Tarski’s theory. Not only does Tarski’s reliance on a specific example leave the precise nature of his general method an open question, but many philosophical issues pertinent to his philosophical goal are not decided by this method. Not surprisingly, the interpretations we will discuss in the present paper involve revision and/or extension of Tarski’s original theory.

2. The Methodological Aim. A secondary yet important goal of Tarski’s theory is to contribute to the methodology of the deductive sciences, or (using Hilbert’s terminology) metamathematics. Although by ‘deductive sciences’ Tarski primarily understood mathematical disciplines presented ‘in the shape of formalized deductive theories’ (Tarski, 1936b, p. 409), most philosophical interpretations of his theory have concentrated on its applications to science and everyday discourse. Tarski’s intended contribution to the methodology of the deductive sciences was both negative and positive. His negative goal is described by Vaught as follows:

[During the 1920s] Tarski has become dissatisfied with the notion of truth as it was being used. Since the notion “σ is true in Il” is highly intuitive (and perfectly clear for any definite σ), it had been possible to go even as far as the completeness theorem by treating truth (consciously or unconsciously) essentially as an undefined notion – one with many obvious properties. . . . But no one had made an analysis of truth, not even of exactly what is involved in treating it in the way just mentioned. At a time when it was quite well understood that ‘all of mathematics’ could be done, say, in ZF, with only the primitive notion e, this meant that the theory of models (and hence much of metalogic) was indeed not part of mathematics. It seems clear that this whole state of affairs was bound to cause a lack of sure-footedness in metalogic . . . . [Tarski’s] major contribution was to show that the notion “σ is true in Il” can simply be defined inside of ordinary mathematics, for example, in ZF. (Vaught, 1974, p. 161)

The positive goal contains both (i) the definition of central methodological (metamathematical) concepts, and (ii) the establishment of methodological (metamathematical) results. Among the methodological definitions and results that Tarski arrived at in the course of his work on truth (late ’20s and early ’30s) are the definition of definability (the notion of ‘object (set) X is definable by expression e’), and the undefinability result (which says that the set of all true sentences of a reasonably rich, bivalent, deductive system S cannot be defined in S). (Tarski, 1931 and 1933, respectively.)

In a later paper Tarski mentioned another methodological goal: ‘bring [the] method [of truth] into harmony with the postulates of the unity of science and of physicalism’. (Tarski, 1936a, p. 406) ‘Unity of science’ and ‘physicalism’ refer to methodological principles advocated by the Vienna Circle, and, as in the case of ‘truth’, Tarski translated these material principles into essentially formal constraints: (i) the definition of truth shall satisfy the requirement of formal rigor, and (ii) the definition of truth shall eliminate all semantic notions (since semantic notions are neither