In *New Essays on Tarski and Philosophy*, ed. by
Douglas Patterson, Oxford: Oxford Univ. Press 2008: 133–57

Tarski's Convention T and the Concept of Truth

Marian David
University of Notre Dame

In this paper, I want to discuss in some detail the original version of Tarski's condition of adequacy for a definition of truth, his *Convention T*. I will suggest that Tarski designed Convention T to serve two functions at once. I will then distinguish two possible interpretations of Tarski's work on truth: a standard interpretation and a non-standard, alternative interpretation. On the former, but not on the latter, the very title of Tarski's famous article about *the* concept of truth harbors a lie.

1. Convention T

Convention T, the original version of Tarski's condition of adequacy for a definition of truth, can be found in his article "The Concept of Truth in Formalized Languages" (Tarski 1983), henceforth CTF. This 1983 English translation revises a 1956 English translation which is based on the German version of the article (Tarski 1935)—compared to the German, the revised English translation adds and/or changes a substantial number of footnotes. The German version, in turn, is an expansion of the Polish original (Tarski 1933), adding an important Postscript.[1]

I quote Convention T as it appears in §3 of CTF, together with a large bit of the paragraph preceding it which helpfully brings up most of the notions and ideas that play an important role in Tarski's thinking leading up to the convention, including ones that did not quite make it into the convention itself (CTF: 187-8):

> Let us try to approach the problem from a quite different angle, by returning to the idea of a semantical definition as in §1. As we know from §2, to every sentence of the calculus of classes there corresponds in the metalanguage not only a name of this sentence of the structural descriptive kind, but also a sentence having the same meaning…In order to make clear the content of the concept of truth in connexion with some one concrete sentence of the language with which we are dealing we…take the scheme ['*x is a true sentence if and only if p*'] and replace the symbol '*x*' in it by the name of the given sentence, and '*p*' by its translation into the metalanguage. All sentences obtained in this way…naturally belong to the metalanguage and explain in a precise way, in accordance with linguistic usage, the meaning of phrases of the form '*x* is a true sentence' which occur in them. Not much more in principle is to be demanded of a general definition of true sentence than that it should satisfy the usual conditions of methodological correctness and include all partial definitions of this type as special cases; that it should be, so to speak, their logical product. At most we can also require that only sentences are to belong to the extension of the defined concept…

[1] In the Postscript, the central theorem of §5 is withdrawn and replaced by the theorem now familiar to logicians as *Tarski's Theorem*, saying roughly that truth is definable for an object-language if, but only if, the metalanguage in which truth is to be defined is "essentially richer" than the object-language; cf. CTF, p. 273; and Tarski's 1944, sec. 10.

Using the symbol '*Tr*' to denote the class of all true sentences, the above postulate can be expressed in the following convention:

CONVENTION T. *A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an* adequate definition of truth *if it has the following consequences:*

(α) all sentences which are obtained from the expression 'x ∈ Tr if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage;

(β) the sentence 'for any x, if x ∈ Tr then x ∈ S' (in other words 'Tr ⊆ S').

The formulation of Tarski's adequacy condition which appears as Convention T at the end of this passage differs in various respects from the formulations he gives in his more popular writings on truth (cf. Tarski 1936, 1944, 1969). It also differs from the formulations that can be found in most discussions of his writings by others. Let us take a closer look at the whole passage.

The language of the calculus of classes to which Tarski refers at the beginning of our passage is completely specified in §2 of CTF. It is a formalized language containing an in-principle inexhaustible stock of variables, '*x,*', '*x,,*', '*x,,,*',…, called the first, second, third, … variable, and only four constants: the universal quantifier, 'Π', the negation sign, '*N*', the sign of alteration (disjunction), '*A*', and the inclusion sign, '*I*'. The variables are interpreted to range over classes of individuals so that, e.g., the sentence 'Π*x,Ix,x,*' says that, for all classes *a*, *a* is included in *a*. You may note that this is a rather meek language, lacking the means to express that an individual or a class is an element of some class. Tarski refers to this language as "the language of the calculus of classes"—as if there were only one. He does not give it a proper name. I will give it a proper name for definiteness: I will call it *Calish*. Distinguish Calish from the calculus of classes. Calish is a formalized language expressing the calculus of classes in Polish notation. The calculus of classes can be expressed in various other languages, using alternative, and maybe more familiar, notations. Calish is the *object-language* for which Tarski will construct a definition of truth in §3 of CTF—but Tarski does not yet use the term 'object-language' in CTF (though he will use it later, in Tarski 1944).

A *structural-descriptive name* of an expression names an expression by spelling it out, without using quotation marks. The following structural-descriptive name—'the expression which consists of five successive expressions, namely the universal quantifier, the first variable, the inclusion sign, the first variable, and the first variable'—names the sentence 'Π*x,Ix,x,*'. The sentence belongs to Tarski's object-language, Calish. Its structural-descriptive name belongs to the *metalanguage* to which Tarski alludes at the beginning of our passage. Tarski usually refers to this language simply as "the metalanguage". He does not give it a proper name. I will call it *Meta-Calish*. This is the language *in* which Tarski will construct his definition of truth *for* Calish. Meta-Calish is also characterized in §2 of CTF—in some (but not in complete) detail. It contains structural-descriptive names of all expressions of Calish. It also contains expressions that allow one to translate all the sentences of Calish into different sentences of Meta-Calish with the same meaning. The sentence 'for all classes *a*, *a* is included in *a*' is such a sentence: it is a translation into Meta-Calish of the Calish sentence 'Π*x,Ix,x,*'. Meta-Calish also contains various other expressions, among them expressions that occur in the

schema quoted in part (α) of Convention T and in the sentence quoted in part (β).

The following sentence,

(1) The expression which consists of five successive expressions, namely the universal quantifier, the first variable, the inclusion sign, the first variable, and the first variable, is a true sentence if and only if, for all classes *a*, *a* is included in *a*,

is an instance of the schema '*x is a true sentence if and only if p*'. The instance is constructed in accordance with the instructions Tarski gives in the paragraph preceding the convention. As he puts it in that paragraph, it is supposed "to make clear the content of the concept of truth in connexion with some one concrete sentence of the language with which we are dealing", namely in connection with the sentence 'Π*x*̣*Ix*̣*x*̣' of Calish. So (1) is one of those sentences he refers to in the paragraph as *partial definitions* which "explain in a precise way, in accordance with linguistic usage, the meaning of the phrase '*x* is a true sentence' which occurs in them". Later, in Tarski (1944: sec. 4), the schema will be labeled 'T', and partial definitions like (1) will be called 'equivalences of the form T'—nowadays, they are often referred to as *T-sentences* or as *T-biconditionals*.[2]

Structural-descriptive names become excruciatingly cumbersome with increasing length.[3] In practice, quotation-mark names are much easier to decode than structural-descriptive names, which is why they are used much more frequently. So, one might alternatively think of (1) in terms of the more perspicuous:

(1*) 'Π*x,Ix,x,*' is a true sentence if and only if, for all classes *a*, *a* is included in *a*.

This is the form in which T-biconditionals are more usually given.[4] Note, however, that part (α) of Convention T refers to structural-descriptive names rather than quotation-mark names. This is because the official metalanguage Tarski has specified in §2 of CTF, Meta-Calish, employs only structural-descriptive names to talk about the expressions of Calish. As far as CTF is concerned, T-biconditional (1) is an official partial definition, whereas T-biconditional (1*) is merely a helpful device for fixing ideas.

Tarski uses the symbols '*Tr*' and '*S*' in his formulation of Convention T. Taking

---

[2] The label 'biconditional' is much better than 'equivalence'. The latter misleadingly suggests a *relational* claim, i.e. a claim of the form '[NAME] is equivalent with [NAME]'. But being a biconditional, i.e. taking the form '[SENTENCE] if and only if [SENTENCE]', (1) does not make a relational claim. Tarski is aware of the potential for confusion on this score (cf. Tarski 1946: chap. 2) and complains about what appears to be an instance of such a confusion in his 1944, section 15. He nevertheless keeps referring to the T-biconditionals as "equivalences" in his 1944 and 1969.

[3] Actually, the official structural-descriptive name of 'Π*x,Ix,x,*' is still more cumbersome than the one I have given. According to CTF, p. 172, it has to be constructed by repeated application, with embeddings, from 'the expression which consists of two successive expressions *x* and *y*'. (If you try to work this out, you will find it difficult to get a grammatical result.) To avoid such complexities, Tarski introduces various symbolic devices for abbreviating structural-descriptive names; e.g. the abbreviated name of 'Π*x,Ix,x,*' looks like this: 'un{(v_1{(in{(v_1{ v_1}))}'.

[4] It is still more usual to give them for cases in which one's object-language is contained in one's metalanguage; e.g. the biconditional

'Π*x,Ix,x,*' is a true sentence if and only if Π*x,Ix,x,*

would be a Meta-Calish T-biconditional, if Meta-Calish did contain Calish, as it does not, and if it did contain quotation-mark names, as it does not.

the second first, note that the expression '$x \in S$' appears in part (β) of the convention without introduction. We are supposed to remember from §2 of CTF that Tarski treats this expression as a notational variant (a symbolic abbreviation) of '*x* is a sentence'—he does this even though '$x \in S$' is short for '*x* is an element of the class of all *x* such that *x* is a sentence'. Both, '*x* is a sentence' and its abbreviation, '$x \in S$', belong to Meta-Calish; they have been defined together to pick out the sentences (closed well-formed formulas) of Calish—at the point where Tarski gives this definition, in §2 of CTF, he has already told us that the expression 'is an element of', together with its symbolic abbreviation, '$\in$', together with various other expressions from general set theory, such as 'the class of all *x* such that', belong to Meta-Calish.[5]

As Tarski indicates in the sentence introducing Convention T, the expression '$x \in Tr$' is short for '*x* is an element of the class of all *x* such that *x* is a true sentence'. He nevertheless treats it as a notational variant of '*x* is a true sentence', and in part (α) of Convention T, he uses it, rather than the more familiar '*x* is a true sentence', to formulate the schema for constructing T-biconditionals such as (1).

When Tarski characterizes his metalanguage in §2 of CTF, he does not list 'true sentence', or '*Tr*', among the expressions belonging to Meta-Calish—not yet. For, as he sees it, the issue of whether truth is definable for Calish within Meta-Calish is the question of whether a logically simple expression, like 'true sentence' or '*Tr*', can be properly introduced into Meta-Calish. Or, to put this somewhat differently, the question is whether Meta-Calish, as characterized in §2, can express the concept of truth for Calish with some combination of expressions already available in it—this is what Convention T is about. If it can, then a simple expression like 'true sentence', or '*Tr*', can be properly introduced into Meta-Calish. So, in a sense, these expressions don't matter: they merely serve as outward signs that the relevant concept is already expressible in Meta-Calish without them. In another sense, however, these expressions do matter, for they serve to remind us of the concept whose definability in Meta-Calish is being investigated.

Tarski's use of '$x \in Tr$' as a variant for '*x* is a true sentence', in combination with part (β) of Convention T, also serves to remind us that he thinks of '*x* is a true sentence' as a logically simple, fused, predicate, along the lines of '*x* is a truesentence', rather than the logically complex predicate '*x* is true and *x* is a sentence'. If he did think of the predicate in the second way, part (β) of Convention T would be entirely superfluous.

Convention T is formulated as a sufficient condition: it says that a definition of '*Tr*' will be called an adequate definition of truth, *if* it has the sentences described in (α)

[5] Tarski defines 'sentence' and '*S*' in one breath in Definition 12 (CTF: 178), saying "*x is a* sentence (*or a* meaningful sentence)*—in symbols* $x \in S$*—if and only if*…". He appears quite unconcerned by what looks to be a difference in ontological commitment between '*x* is a sentence' and '*x* is an element of the class of all sentences'. There is a similarly unconcerned passage in his *Introduction to Logic*, where he says that any sentential function with one free variable can be transformed into an equivalent function of the form '$x \in K$', where in place of '*K*' we have a constant denoting a class, so that one may "consider the latter formula as the most general form of a sentential function with one free variable" (Tarski 1946: 70-1). By the way, there is nothing in the definiens of Definition 12, or in the definiens of the definiens, corresponding to the parenthetical 'meaningful': the definition of 'sentence' is given "by means of purely structural [i.e. syntactic] properties" (CTF: 166). The parenthesis is merely a reminder that the expressions to which the syntactically defined term 'sentence' applies, i.e. the well-formed formulas of the formalized language Calish, are assumed to have their ordinary meanings: "We shall always ascribe quite concrete and, for us, intelligible meanings to the signs which occur in the languages we shall consider" (CTF: 167).

and the sentence mentioned in (β) as consequences; it does *not* say *only if*. This may come as a surprise, because Tarski's adequacy condition is almost always presented as a sufficient *and necessary* condition. Moreover, at other places Tarski himself puts it as a sufficient and necessary condition; e.g. the first time he formulates it in Tarski 1944 (sec. 4).[6] What are we to make of this? There are two clear indications in CTF that Tarski intends Convention T as a sufficient *and* necessary condition. First, in the text preceding the convention, he says, commenting on the relevant instances of the schema '*x is a true sentence if and only if p*', that "not much more in principle is to be demanded of a general definition of true sentence than that it should…include all partial definitions of this type as special cases…". If this much *is to be demanded*, then having the sentences described in (α) as consequences is intended as a necessary condition. He goes on to say: "At most we can also require that only sentences are to belong to the extension of the defined concept…". So, having the sentence mentioned in (β) as consequence is also intended as a necessary condition. Second, in a footnote appended to Convention T, Tarski says that "after unimportant modifications" of its formulation, the convention would "become a normal definition" (CTF: 188)—and a normal definition requires an 'if and only if'. Tarski's later book, *Introduction to Logic*, provides additional, circumstantial evidence. There he points out that mathematicians prefer the word 'if' when laying down definitions: "what we have here", he says, "is a tacit convention to the effect that "if" or "in case that", when used to join definiendum and definiens, are to mean the same as the phrase "if, and only if" ordinarily does" (Tarski 1946: 36). As far as Convention T as stated in CTF is concerned, there seems to be sufficient evidence that, despite first appearances to the contrary, it is intended as a necessary as well as a sufficient condition.[7]

Convention T is focused on the "adequacy", or "material adequacy" (CTF: 186), of a definition of truth. It takes for granted that the definition whose material adequacy is under consideration is *formally correct*. Tarski tells us surprisingly little about this presupposed condition of formal correctness. In the paragraph preceding Convention T, he alludes briefly to "the usual conditions of methodological correctness", but he does not indicate what they are. He does say, early on in CTF, that a question about the definability of a concept "is correctly formulated" only if one gives a list of the terms one intends to use in constructing the definition, that these terms "must admit of no doubt", and that he will not make use of any semantic concept in his definition if he is "not able previously to reduce it to other concepts" (CTF: 152-3). Later, in §2 of CTF, he lists the expressions belonging to his metalanguage, Meta-Calish under two general headings: expressions of "a general logical character" and expressions of "a structural-descriptive character" (CTF: 169-73). Though it would surely have been pertinent at this point, he does not remind us of the condition of formal correctness: he does not bother to emphasize that the vocabulary of Meta-Calish, which he has just specified, does not contain any undefined or unreduced semantic expressions. Quite a bit later, when commenting on (the original version of) his negative theorem in §5, he describes his

[6] However, the second and more official formulation in section 4 of Tarski's 1944 states it only as a sufficient condition. But then again, when he briefly restates it in section 9, he specifically reminds us of the necessity of the condition.

[7] For more discussion of this topic see Patterson (2006) who, looking also at later versions of Tarski's condition of adequacy, arrives at a somewhat different conclusion.

article as an attempt to "reduce" semantic concepts to structural-descriptive, i.e. syntactic, concepts (CTF: 252). He points out that the attempt fails in the end. Though the reduction succeeds with respect to certain "poor" object-languages, such as the language of the calculus of classes, i.e. Calish, it does not go through with respect to "rich" object-languages, such as the language of the general theory of classes, i.e. the language of general set theory (CTF: 253-4). The reduction fails in cases of the latter sort because, as the theorem tells us, it is impossible in such cases to give an adequate definition of truth on the basis of the metatheory, *if* the metatheory is consistent (CTF: 247). In other words, the envisaged definition itself or, more generally, the metatheory to which the definition would belong, would be inconsistent. As far as I can see, this failure seems to concern the condition of formal correctness rather than material adequacy, or if it concerns material adequacy then only because material adequacy presupposes formal correctness according to Convention T. Tarski does not comment on this—though it is noteworthy that he tends to talk in terms of 'correctness' rather than 'adequacy' when reflecting on the import of his negative theorem.

In §2 of CTF, Tarski distinguishes between the meta*language*, on the one hand, and the meta*theory*, on the other (CTF: 167). The latter contains a system of axioms and definitions formulated in the metalanguage. For example, the definitions of sentencehood and truth for Calish belong to the metatheory for Calish formulated in Meta-Calish. What it the metatheory a theory of? At bottom, it is an axiomatic theory of the syntactic structure of the object-language: "What we call the metatheory is, fundamentally, the *morphology of language*—a science of the form of expressions" (CTF: 251). Its axioms do not contain any semantic notions. Such notions appear only in the definitions and then only if they are ultimately defined in terms of (reduced to) non-semantic notions. The *axioms* of the metatheory play a rather important though strangely unacknowledged role in Convention T. Before we get to this point, let us take a closer look at these axioms.

There are two groups of axioms in the metatheory: one group Tarski calls "the general logical axioms"; the other one he calls, somewhat awkwardly, "the specific axioms of the metalanguage" (CTF: 173). Tarski does not list any examples from the first group, the one he *calls* "the general logical axioms". He merely says that they are well-known (referring us to Whitehead and Russell's *Principia Mathematica*) and that they "suffice for a sufficiently comprehensive system of mathematical logic" (CTF: 173). It is clear, however, that in addition to typical logical axioms this group is also supposed to contain the axioms of general set theory. Since general set theory is *not* regarded as belonging to logic nowadays, this makes Tarski's label for this group of axioms rather problematic from our present point of view. But there is no deliberate misdirection involved here. At the time when Tarski composed the original version of CTF, it was fairly widely held that general set theory does belong to logic; and Tarski evidently held the view too—one may note that the inventory of his metalanguage lists expressions such as 'is included in', 'is an element of', 'class', 'infinite class', 'ordered pair', 'sequence', and 'natural number' under the heading "expressions of a general logical character" (cf. CTF: 170-1). One may note also that about a decade later Tarski was already rather more skeptical about the logical nature of set theory.[8]

[8] In his 1944, endnote 12, Tarski first points out that he is using 'logical' in "a broad sense" in which it comprehends "the whole theory of classes and relations (i.e., the mathematical theory of sets)". But he then remarks that he is "personally inclined" to use the term in "a much narrower sense, so as to apply it only to

Tarski does provide a list of the axioms of the second group, the one he calls "the specific axioms of the metalanguage". They describe syntactic properties and relations of his object-language, Calish. The first four are mostly concerned with identity conditions of simple and complex expressions of Calish. Axiom 1, for example, says that the negation sign, the alternation sign, the sign for the universal quantifier, and the inclusion sign "are expressions, no two of which are identical" (CTF: 173). The upshot is that claims about the non-identity of intuitively different expressions of Calish become enshrined as axioms in the Meta-Calish metatheory. Note that non-identity claims, syntactic or otherwise, are not logical truths and that Tarski does not regard these syntactic non-identity claims as logical truths, even though he lays them down as axioms: he clearly separates the axioms from this second group from the ones he refers to, problematically, as the general logical axioms. Note also that Tarski's labels for the two groups of axioms suggest that he had the following picture in mind. The axioms of the first group, including the axioms of set theory, belong in one form or another to any metatheory, independently of the specifics of the object-language and metalanguage at hand. The axioms of the second group, on the other hand, will depend entirely on the specifics of the metalanguage at hand, that is, ultimately on the specifics of the object-language to which the metalanguage belongs, because these axioms specify the syntax of the object-language in terms of its metalanguage.

Let us return now to Convention T. In the paragraph preceding the convention, Tarski says that an adequate definition of truth should *include* as special cases all partial definitions, that is, all the relevant T-biconditionals such as (1). In Convention T itself, Tarski does not use the term 'include', he uses the term 'consequence'. Specifically, he says that a definition of '*Tr*' will be called an adequate definition of truth if it has as *consequences* the sentences described in part (α) of the convention, i.e. the relevant T-biconditionals, as well as the sentence mentioned in part (β). Tarski does not mention the axioms of the metatheory here and his wording creates the impression that he means all these sentences to be consequences of the definition taken just by itself.

This is curious. For when one looks at the definition Tarski constructs in §3 of CTF, which he says is an adequate definition of truth, it turns out that it does not by itself have the T-biconditionals as consequences. At least, it does not have them as formal consequences—they are not *derivable* from the definition of truth alone: the axioms (and more definitions) of the metatheory are needed as additional premises. Moreover, at times Tarski *does* mention the metatheory or its axioms in this connection. When he describes the problem of constructing a definition that satisfies Convention T, later in CTF, he says that "it is a question of *whether on the basis of the metatheory of the language we are considering the construction of a correct definition of truth in the sense of convention T is in principle possible*" (CTF: 246); and when he states a version of his adequacy condition in the short paper "The Establishment of Scientific Semantics", he talks of the "provability" of the T-biconditionals "on the basis of the axioms and rules of inference of the metalanguage" (Tarski 1936: 404). Still, in Convention T itself, Tarski suppresses any reference to the axioms of the metatheory. Why he does this is unclear, and it seems fair to say that the wording of Convention T is rather misleading on this point.[9]

---

what is sometimes called "elementary logic," i.e., to the sentential calculus and the (restricted) predicate calculus".

[9] In the various versions of his adequacy condition which Tarski gives in his later papers, he makes again

One might suggest that Tarski may have had in mind some relation other than derivability when using the term 'consequence' in Convention T. This seems unlikely: he would have said so. Moreover, note that in §2 of CTF he does define 'consequence' for formulas of Calish so that it refers to a formal/syntactic derivability relation (CTF: 182). Admittedly, the term 'consequence' thus defined in §2 applies to a relation between items of Tarski's object-language, Calish, whereas the term 'consequence' that appears in Convention T applies to a relation between items of Tarski's metalanguage, Meta-Calish. Still, it would be very strange if Tarski had reused the same term in Convention T while having something fundamentally different in mind.

One might suggest that Tarski in Convention T thought it alright to suppress references to the additional premises that are needed for deriving the T-biconditionals from an adequate definition of truth because these additional premises consist of *axioms*. This may be so; he does not say. But consider that one would normally regard this as acceptable only if the axioms involved express logical truths. As I have pointed out, Tarski at the time subsumed set theory under logic. This may make it understandable why Convention T does not mention any need for set-theoretic axioms when talking about the consequences of an adequate definition of truth. But Tarski did not subsume the principles that describe the syntax of an object-language under logic. This makes it difficult to understand why Convention T does not mention the need for the syntactic axioms.[10]

The issue is of some interest because it bears on how one should conceive of an adequate definition of truth in relation to the T-biconditionals. In the paragraph preceding Convention T, and at various other places in CTF, Tarski describes a definition of truth that is adequate in the sense of Convention T as the "logical product" of the T-biconditionals and refers to the latter accordingly as "partial definitions".[11] But, as we have seen, among the axioms needed for deriving the T-biconditionals there are axioms that are not logical in nature, not even by Tarski's own lights at the time he composed CTF. It is then hard to see how Tarski can avoid the censure that it is both misleading and wrong to describe a definition of truth that is adequate in the sense of Convention T as

---

no reference to the axioms of the metatheory. He says that an adequate definition "implies" the T-biconditionals, that the T-biconditionals "follow from" an adequate definition (Tarski 1944: sec. 4), that an adequate definition "enables us to ascertain" the T-biconditionals (Tarski 1969: 106), and that such a definition will "imply" the T-biconditionals "as consequences" (Tarski 1969: 114).

[10] Set-theoretic axioms in the metatheory are needed for deriving the relevant T-biconditionals from the definition of truth only if the object-language under consideration is similar in complexity to Calish; they are not needed for object-languages with finitely many sentences. Syntactic non-identity axioms, however, are needed even for finite object-languages (as long as they contain more than one sentence). Take an object-language with only two sentences, say 's_1' and 's_2', and assume for simplicity's sake that the meta-language contains the object-language and contains quotation-mark names, ''s_1'' and ''s_2'', of the two sentences of the object-language. Tarski tells us that the following will be an adequate definition of truth for this language (cf. CTF: 188):

$x \in Tr$ if and only if (x = 's_1' and s_1) or (x = 's_2' and s_2).

Deriving the T-biconditional with respect to sentence 's_1', i.e.,

's_1' $\in$ *Tr* if and only if s_1,

requires the premises ''s_1'' = 's_1'' and ''s_1' $\neq$ 's_2''. The former is an instance of a logical truth; it will thus be covered by some non-problematic member of the group of "general logical axioms". The second is not an instance of a logical truth; it would have to be laid down as a (the sole) syntactic non-identity axiom for this particular object-language.

[11] See, e.g., CTF, pp. 155, 157, 163, 165, 187, 236, 238, 253.

the *logical* product of the T-biconditionals.[12]

2. The Double-Life of Convention T

I think that Tarski wants Convention T to play a double-role, to serve two important functions at once. On the one hand, and taken strictly, it is supposed to provide a necessary and sufficient condition of adequacy for a definition of truth for Tarski's object-language, Calish, formulated in Tarski's metalanguage, Meta-Calish. In this role the condition is radically non-general. On the other hand, and taken more loosely, the condition is also supposed to function in the role of an exemplar, to intimate or convey something rather more general than it strictly speaking expresses.

Imagine you first encountered Convention T all by itself. (You may want to reread the convention at this point but without the preceding paragraph.) You would then still understand it in a rough way, but you would also notice that the crucial terms 'the metalanguage' and 'the language in question' have strangely floating or indefinite or indeterminate reference. One of the first questions coming to your mind ought to be: "Which metalanguage?" and "What is meant by *the* language in question?" These questions receive no answer from within Convention T. When, on the other hand, Convention T is seen in the context of CTF, the references of these terms becomes clear. The term 'the language in question' refers back to Tarski's object-language, Calish, the language of the calculus of classes, which he mentions explicitly early in the preceding paragraph and which he has specified in §2 of CTF. The term 'the metalanguage' picks up on his allusion to *the* metalanguage, also early in the preceding paragraph, which allusion in turn refers back to Meta-Calish, his metalanguage, which he has characterized in §2 of CTF.

Remember in addition the following three points: the symbol '*S*', which appears without introduction in part (β) of Convention T, was defined earlier, in §2 of CTF, to pick out the sentences of Calish; the sentence quoted in part (β) belongs to Meta-Calish; and the schema for the T-biconditionals, quoted in part (α), also contains material belonging to Meta-Calish.[13]

In sum, Convention T, taken strictly, is maximally specific. Moreover, it is maximally specific along two different dimension. The first one is more frequently acknowledged: strictly speaking, Convention T talks only about the adequacy of a truth definition for the one object-language Calish. Recognizing this, one might still think that Convention T is at least general along another dimension: one might still think that it gives a condition of adequacy for truth definitions for Calish in arbitrary metalanguages. But no, it doesn't, for it refers back to, and contains quoted material from, the one metalanguage Meta-Calish.[14]

This second dimension of specificity can be illustrated by comparing Convention T with Konvention W from the German version of Tarski's article (Tarski 1935: 305-6). Tarski's object-language in the German version is the same as in CTF, Calish: it is

[12] I should point out though that Tarski tends to qualify his talk of the definition as a logical product with a "so to speak", cf. CTF, pp. 187 and 238.

[13] Though, as far as I can tell, the '*p*' in the schema does not belong to Meta-Calish. Tarski lists a '*p*' as belonging to his metalanguage on p. 173 of CTF, but since it is said to represent (a sequence of) natural numbers, this can't be the '*p*' that appears in the schema.

[14] Gupta and Belnap mention this point briefly in their 1993, p. 2, footnote 4.

specified as containing the same four constants of Polish notation and the same in-principle inexhaustible stock of variables. But Tarski's metalanguage in the German version is not the same as the one in CTF. Take the Calish sentence '$\Pi x_{\circ} I x_{\circ} x_{\circ}$'. The metalanguage specified in §2 of the German version, call it *Meta-Kalisch*, does not contain the structural-descriptive name 'the expression which consists of…' (etc.), nor does it contain the sentence 'for all classes *a*, *a* is included in *a*'. Instead, it names the sentence '$\Pi x_{\circ} I x_{\circ} x_{\circ}$' with the structural-descriptive name 'der Ausdruck, der…' (etc.) and translates the sentence as 'für jede beliebige Klasse *a*, *a* ist in *a* enthalten'. Moreover, in place of Meta-Calish expressions such as 'if and only if', 'is an element of' (or '$\in$'), 'class of all *x* such that', and many others, Meta-Kalisch contains 'dann und nur dann, wenn', 'ist ein Element von' (or 'ε'), 'Klasse aller solchen *x*, dass', and many others. The vocabularies of the two metalanguages are almost completely different.

This has the following consequences. Convention T and Konvention W state different conditions. That's because Convention T quotes expressions belonging to Meta-Calish but not to Meta-Kalisch and Konvention W quotes expressions belonging to Meta-Kalisch but not to Meta-Calish—part (α) of Konvention W mentions the schema '*x ε Wr dann und nur dann, wenn p*' and part (β) mentions the sentence '*für ein beliebiges x, wenn x ε Wr, so x ε As*'. And this means, furthermore, that a definition of truth for Calish that is adequate by the lights of Convention T will not be judged adequate by the lights of Konvention W, and vice versa. Take the definition of truth for sentences of Calish Tarski constructs in §3 of CTF. It is formulated in Meta-Calish. Assume that it is an adequate definition according to Convention T, as Tarski maintains. It then has among its consequences all the Meta-Calish sentences constructed from the schema '*x ∈ Tr if and only if p*' by following the instructions given in part (α) of Convention T, as well as the one Meta-Calish sentence quoted in part (β) of the convention. However, since Meta-Calish does not contain the relevant expressions of Meta-Kalisch, the Meta-Calish definition does *not* have as consequences the sentences of Meta-Kalisch referred to in parts (α) and (β) of Konvention W: these sentences cannot even be formulated in Meta-Calish. This goes the other way round too. In §3 of the German version of CTF, Tarski constructs a definition of truth for Calish in Meta-Kalisch. Assume that it is an adequate definition according to Konvention W, as Tarski maintains. It then has among its consequences all the Meta-Kalisch sentences constructed from the schema '*x ε Wr dann und nur dann, wenn p*' by following the instructions given in part (α) of Konvention W, as well as the one Meta-Kalisch sentence quoted in part (β) of Konvention W. But since Meta-Kalisch does not even contain the relevant expressions of Meta-Calish, the definition given in the German version of CTF does *not* have as consequences the sentences of Meta-Calish referred to in parts (α) and (β) of Convention T. Note that there is no conflict here: Convention T merely implies that the German version's definition of truth for Calish, which is constructed in Meta-Kalisch, is not an adequate definition of truth for Calish in Meta-Calish (and vice versa), which is obvious enough. But the situation illustrates how radically specific Convention T really is.

If Convention T is thus radically specific, Why didn't Tarski make this more explicit? His condition of adequacy is surely one of the centerpieces of his article, Why did he use the indeterminate terms 'the metalanguage' and 'the language in question' whose precise reference is determined only through the larger context in which the condition occurs? This has to do with the other, more elusive role Convention T is

supposed to play.

In the second sentence of CTF Tarski signals that he is going to give a definition of truth only for one individual language: when announcing that it is the task of his article to construct a materially adequate and formally correct definition of the term 'true sentence', he inserts the rather important qualification "with reference to a given language" (CTF: 152). But later, in §2—after he has abandoned the attempt to give a definition of truth for ordinary languages and has announced that he is going to restrict his attention to formalized languages and has drawn the object-language/metalanguage distinction, but just before he begins to lay down his particular object-language and his particular metalanguage—he remarks that it is possible to give a method for defining truth "for an extensive group of formalized languages" (CTF: 167). He then says that giving a general description of this method and of the languages for which the method works "would be troublesome and not at all perspicuous" (CTF: 168), and that he therefore prefers to introduce us to this method in another way:

> I shall construct a definition of truth of this kind in connexion 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. (CTF: 168)[15]

It is this strategy—using a particular exemplar to convey something more general—which is, I believe, the reason that leads Tarski to formulate Convention T as he does, with the referentially indeterminate terms 'the metalanguage' and 'the language in question'. On the one hand, the actual context in which the convention occurs *does* provide the proper determinate references for Tarski's use of these terms in the convention, namely Meta-Calish and Calish respectively, while on the other hand, the indeterminate terms also impart a suggestion of generality to the convention. (Remember that one can uses phrases like 'the dog…' to refer to a particular, contextually salient dog but that one can also use them to express generalizations about all, or all typical dogs.) Strictly speaking, this suggestion of generality is of course wrong: Convention T does not really talk about truth definitions for arbitrary object-languages, not even about truth definitions for some range of object-languages that are of a logical construction similar to Calish, nor does it talk about truth definitions for Calish in arbitrary metalanguages or in a range of metalanguages similar to Meta-Calish. Nevertheless, the suggestion of generality carried by the indeterminate terms manages to impart the desired message, namely that, by taking Convention T as our model, we could relatively easily formulate "analogous" adequacy conditions for other object-language/metalanguage pairs. The message comes through even though we—and I believe Tarski too—would be very hard pressed to spell out what 'analogous' actually amounts to here.[16]

---

[15] Compare also the Introduction of CTF, where he says that "there is a uniform method of construction of the required definition" for each of the "poorer" languages for which truth is definable and announces that he will "carry out this construction for a concrete language in full and in this way facilitate the general description of the above method which is sketched in §4" (CTF: 153-4). How "uniform" the method of construction actually is is a difficult question.

[16] What would it take to formulate a free-standing, context independent, general condition of adequacy for arbitrary object- and metalanguages? At the very least it would require finding a formulation that does not

The use of referentially indeterminate terms in this double-role is in fact a pervasive feature of CTF: it can be found in Tarski's use of the most basic vocabulary of his metalanguage. When he first introduces Calish, on p. 168 of CTF, he says it contains the universal quantifier, 'Π', (together with some other signs). Six pages later, when he introduces the structural-descriptive terms of Meta-Calish, he lists (among many others) the term 'the sign of the universal quantifier'. This is a referentially indeterminate term: it could refer to other signs, e.g. to '∀', or to Russell's '( )'. But Tarski takes for granted that the reference of his uses of 'the sign of the universal quantifier' throughout CTF has been fixed by the initial context on p. 168 where he set up Calish to contain 'Π', referring to it as *the* universal quantifier. Note how this allows this indeterminate term to play its second, generality-intimating role. If you wanted to construct a definition of truth, not for Calish but for an object-language that expresses the calculus of classes in a different way, say by using '∀', you would, as it were, only have to go to p. 168 of CTF, where Tarski introduced his Calish, and replace 'Π' with '∀', thereby referring to '∀' as *the* universal quantifier. You could then introduce and use the term 'the sign of the universal quantifier', which would now belong to your new metalanguage, in the same way in which Tarski used it: following his method, setting up your metalanguage much like he did his, the indeterminate term would now refer back to '∀', rather than 'Π', when used by you in your context.

Consider also Tarski's use of the term 'sentence'. In §2 of CTF, he defines the Meta-Calish term '*x* is a sentence', or '$x \in S$', as the limiting case of a sentential function, one without free variables. The definition, Definition 12 (CTF: 178), invokes the previously defined term, '*x* is a sentential function', whose definition uses (abbreviations of) the referentially indeterminate terms 'the sign of the universal quantifier', 'the negation sign', 'the alteration sign', and 'the inclusion sign', which in Tarski's context refer to the basic vocabulary of Calish, i.e. to 'Π', '*N*', '*A*', and '*I*', respectively (cf. Def. 10, p. 177 of CTF). So Definition 12, even though its definiendum looks like this, '*x* is a sentence' (or '$x \in S$'), defines this term so as to pick out only the sentences (the closed well-formed formulas) of Calish. Nevertheless, when he gives a brief preview of what he is about to do, Tarski announces that he will "obtain the concept of *sentence*" as a special case of the notion of a sentential function (CTF: 176). A natural reaction to this remark would be: "Well, not really, for you are really going to obtain a much more restricted concept, something like *sentence of Calish*". This reaction may well be justified. But note that Tarski does not, and does not seem to want to, talk in these terms. He does not phrase the definiendum of Definition 12 as '*x* is a sentence of the language of the calculus of classes'. He wants the definition to play the generality-intimating role—he himself will go on to use the term 'sentence', without qualifications, to refer to the sentences of Calish (primarily in §3) and later to refer to the sentences of various other languages (in §§4 and 5). The wording of Tarski's definition does not require any modifications, even when employed by others working with other object-languages: in their contexts, the definition will end up picking out the sentences of their object-languages.[17]

---

invoke expressions of Meta-Calish or of any other particular metalanguage. Readers may want to try their hands on this.

[17] This also works, albeit in a rather more limited fashion, when it comes to Definition 10, which defines '*x* is a sentential function'. Since this definition relies on the referentially indeterminate terms 'the inclusion sign', 'the negation sign', etc., the wording of the definition can remain, even if one has earlier introduced

Consider also Tarski's use of the terms 'true sentence' and 'truth'. As I said above, Tarski signals right at the beginning of CTF that he is only going to define 'true sentence' with reference to a given language. He reminds us of this at the beginning of §3. But then, in the remainder of §3, he talks as if this qualifying restriction simply were not there, mentioning it again only once more, at the very end of §3 (cf. CTF: 208). Remember the sentence that immediately precedes Convention T. Tarski says there that he is going to use the symbol '*Tr*', the symbol he is about to mention in the formulation of the convention, "to denote the class of all true sentences". *All* true sentences? Not really: only the true sentences of Calish. Tarski must be using 'all' or 'true sentence' (or both) in a seriously restricted way.[18] Note also the use of the term 'truth' in Convention T: as Tarski formulates it, the convention looks like it specified the condition under which a formally correct definition of '*Tr*' is an adequate definition of *truth*, period. Tarski does not say "an adequate definition of truth for Calish" or something like that: he just says "an adequate definition of truth".

The definition Tarski finally states in §3 of CTF, Definition 23, defines truth only for his specific object-language, Calish—this for reasons analogous to the reasons why his definition of 'sentence' is restricted to the sentences of Calish. Definition 23 defines "*x is a* true sentence—*in symbols* $x \in Tr$—*if and only if* $x \in S$ *and every infinite sequence of classes satisfies x*" (CTF: 195). The definiens presupposes the term 'sequence *f* satisfies the sentential function *x*', and the definition of this term, Definition 22, appeals to the basic vocabulary of Calish (CTF: 193-5). Hence, this term, and consequently the term 'true sentence', or '*Tr*', is defined only for Calish. Nevertheless, when Tarski talks about his definition of 'true sentence', no restriction to Calish is mentioned. In the paragraph before Definition 23, where he gives a brief explanation of what he is about to do, he simply says that "the concept of truth" will be reached on the basis of Definition 22; and right after he has stated his definition, he claims that it is an "adequate definition of truth in the sense of convention T" (CTF: 194, 195). No restriction to his object-language, Calish, is mentioned at all.

Again, it seems he proceeds in this manner because he wants both Convention T and his definition of 'true sentence', or '*Tr*', to play the double-role: on the one hand, and taken strictly, they are maximally specific, playing their role in the project of defining 'true sentence' for Calish within Meta-Calish, while on the other hand, they are supposed to serve as exemplars and are thus formulated without explicit references to Calish and Meta-Calish so as to convey or suggest a more general message that would be rather difficult to state explicitly.

3. True in L?

Standard discussions of Tarski's work on truth present Tarski as having defined a term of the form 'true in L', or a concept *truth in L*, for one specific object-language, Calish, *and*

---

an object-language containing, say, '$\subseteq$' and '$\sim$' instead of Calish which contains '*I*' and '*N*'. But changes in the wording of Definition 10 will become necessary, if one's object-language has a different grammar than Calish, or if it is not a language talking about classes at all so that the term 'the inclusion sign' becomes inappropriate.

[18] In the paragraph preceding the convention he considers what is to be demanded "of a general definition of true sentence [sic]". He means a "general" definition in the sense that it concerns all sentences of Calish, as opposed to the "partial definitions" he has just been talking about, each one of which concerns only one sentence of Calish.

as having shown how to define terms of the form 'true in L', or concepts of *truth in L*, for a range of (well-behaved) object-languages.[19] Convention T is then presented accordingly to state a condition under which a definition, formulated in a metalanguage, of a term of the form 'true in L' is an adequate definition of *truth in L*. Note that you must *not* think here of the 'L' as a genuine variable so that '*x* is true in L' would express a relation holding between sentences and languages—Tarski did not define anything properly expressible by '*x* is true in *y*', with variable '*y*' ranging over languages. Instead, you must think of 'L' as a dummy letter, so that '*x* is true in L' is a schematic way of hinting at various one-place predicates, '*x* is true in ___ ', where a name naming some object-language goes into the gap, e.g. '*x* is true in Calish'.[20]

Presenting Tarski in this way implies that he has not defined the concept *truth*, or *true sentence*, but a different and much more restricted concept, which we might want to call '*truth in Calish*', but that he has also given us guidelines for defining additional such concepts—concepts we can name only after we have named the object-languages we are interested in. For example, we could specify certain (well-behaved) fragments of English and German, name them 'E_0', 'E_1', …, 'G_0', 'G_1', …, and follow Tarski's guidelines to define concepts which we might call '*truth in* E_0', '*truth in* G_0', and so on.

A troubling question arises: What do all these concepts have in common that justifies our using the word 'truth' or 'true' when naming or expressing them? One's first inclination is to respond that these concepts have the following feature in common: the definitions of the terms expressing them, the definitions of '*x* is true in Calish' '*x* is true in E_0', '*x* is true in G_0', etc., all satisfy Convention T. But that can't be quite right, for Convention T is about a definition of '*x* is a true sentence', or '$x \in Tr$', for Calish, formulated within Meta-Calish: it says nothing about '*x* is true in E_0' or '*x* is true in G_0'. Taking Convention T as a model and using Tarski's general guidelines (given in §4 of CTF) for specifying metalanguages for object-languages of certain types, we may be able to formulate an "analogous" convention covering '*x* is true in E_0', and an "analogous" convention covering '*x* is true in G_0', and so on. But these conventions will all be different and different from Convention T, which makes it difficult to spell out how precisely they help answering the troubling question.[21]

I will not pursue this troubling question. Instead, I want to raise an interpretive question, though I won't be able to resolve it: Did Tarski himself hold the view which the standard way of presenting him has him advocate as a matter of course? Interestingly, that's not at all easy to tell, because Tarski himself does not actually use truth-terms with built-in language parameters, terms of the form 'true in L'.

---

[19] Though not, of course, for arbitrary object-languages: that, according to Tarski's Theorem, cannot be done consistently, since it would allow defining a term 'true in L' for the language indicated by 'L' itself, which would lead into paradox.

[20] Note that the 'in' in 'true in L' is very different from the 'in' in 'is defined in a metalanguage'. Note also that the standard way of presenting Tarski in terms of 'true in L' swallows the word 'sentence' that appeared in Tarski's 'true sentence'—remember Tarski's '*Tr*'.

[21] The basic point was well-raised by Quine (1951: 32-7), albeit with respect to Carnap's attempt at explicating *analyticity*. Adapted to Tarski-style definitions, Quine's objection would go like this: Tarski shows us how to define '*x* is true in Calish', '*x* is true in E_0', '*x* is true in G_0', and so on, but not the general notion of truth, not '*x* is true in *y*', with variable '*x*' and '*y*'. The newly defined term '*x* is true in Calish', or rather '*x*-is-true-in-Calish', "might better be written untendentiously as '*K*' so as not to seem to throw light on the interesting word ['true']" (Quine 1951: 33). Compare my 1996, where I ask why Quine did not raise this objection against the idea that Tarski-style definitions throw any light on truth.

To get a better idea of what is involved here, consider a formulation such as

d is an adequate definition of truth for Calish (within the metalanguage Meta-Calish)

This can be parsed in two ways: (*a*) as saying that *d* adequately defines *truth-for/in-Calish* within Meta-Calish; or (*b*) as saying that *d* adequately defines *truth* and does so for Calish and within Meta-Calish.

The standard interpretation of CTF opts for (*a*) on the following grounds. Tarski has not defined '*x* is true in *y*' with variable '*y*' ranging over different object-languages. Instead, he has given a recipe for defining a one-place predicate, '*x* is a true sentence', or '$x \in Tr$', for a range of different object-languages. His recipe involves constructing different definitions which appeal to the expressions of their respective object-languages and assign different extensions to their respective occurrences of the definiendum '*x* is a true sentence'—after all, different object-languages contain different sentences. But this means, so the standard interpretation, that *different concepts* are being defined which, to avoid ambiguity, are best expressed by different definienda, different so-called "truth-predicates" with different built-in language parameters, such as '*x* is true in Calish' and '*x* is true in E_0'.

The alternative, non-standard, interpretation I want to consider here opts for (*b*) in light of the way Tarski expresses himself throughout CTF. He uses 'true sentence', rather than some term or terms of the form 'true in L', and he confidently talks about *the* concept of truth, implying that there is only one: Tarski's way of talking gives few indications that he held a view according to which there are somehow different "truth-concepts" for different languages.

Tarski does, of course, indicate in a number of places that his definition is restricted somehow to his individual object-language, Calish, but not in a way suggestive of the standard 'true in L' interpretation. At the very beginning of the Introduction of CTF, he says his task is "to construct—with reference to a given language—*a materially adequate and formally correct definition of 'true sentence'*" (CTF: 152). At the beginning of §3, he says that he is about to turn to "the construction of the definition of true sentence, the language of the calculus of classes still being the object of investigation" (CTF: 186). At the very end of §3, he says that he has succeeded "*in doing for the language of the calculus of classes what we tried in vain to do for colloquial language: namely to construct a formally correct and materially adequate semantical definition of the expression 'true sentence'*" (CTF: 208). Note the absence of truth terms with built-in language parameters. With hindsight, one might even think that Tarski goes to some lengths to avoid formulations of the form 'true in L'—and this holds not only for CTF but also for his later writings on the subject (cf. Tarski 1944, 1969).[22]

[22] Tarski uses an 'in L' formulation in "Truth and Proof"—but not to talk of 'true in L'. Instead, he uses it to restrict the domain of the quantifier in a definition of 'true': "For every sentence *x* (in the language *L*), *x* is true if and only if…" (Tarski 1969: 106-7). Searching for a source of the contemporary custom of talking in terms of 'true in L', I find Carnap talking of the definition of 'true in *S*' in his *Introduction to Semantics*. However, Carnap's '*S*' is supposed to indicate a "semantical system"—a system of rules, formulated in a metalanguage, containing the syntactic "rules of formation" of a language as well as "rules of designation" and "rules of truth" (Carnap 1942: 22-5). Early in the book, Carnap distinguishes between *languages*, which at first appear to be syntactically individuated, and *semantic systems* of languages: it seems that it

The standard interpretation relies on the principle: *if different extensions, then different concepts*—applied to our case: if different definitions constructed in accordance with Tarski's recipe assign different extensions to different occurrences of 'true sentence', then they define different concepts; that is, they assign different concepts to their respective occurrences of 'true sentence', so that their respective definienda are less ambiguously expressed by different terms such as 'true in Calish', 'true in E_0', etc. The standard interpretation, one might say, assumes that option (*b*) boils down to option (*a*): *defining* truth *for* Calish or *for* E_0 amounts to defining *truth in Calish* or *truth in* E_0.[23]

The standard interpretation has to explain why Tarski's way of expressing himself throughout CTF does not make it at all apparent that he is telling us how to define a range of different "truth-concepts"; it has to explain why he systematically refrains from using terms of the form 'true in L'. To explain this, one may cite Tarski's strategy of proceeding by exemplar: Tarski's Convention T is really about *truth in Calish*, and Tarski really defines the concept *truth in Calish*, within Meta-Calish, but he suppresses explicit references to Calish and Meta-Calish so as to convey a more general recipe for defining a range of different concepts—convey it by means of formulations that appear general even though in the context in which he uses them they really refer to specifics.

On the alternative interpretation, the different definitions of 'true sentence' that can be constructed in accordance with Tarski's recipe are concerned with one concept, *truth*, or better *true sentence*, but they define this one concept *for* different object-languages. The alternative interpretation, one might say, assumes that option (*b*) can stand on its own and does not boil down to option (*a*). This seems feasible only if the transition 'different extensions → different concepts' is not generally reliable—as indeed it isn't: where *context sensitivity* comes into play the transition appears to fail. A context-sensitive term such as 'today' may plausibly be said to express one and the same concept even though different occurrences of the term have different extensions. So, on the alternative interpretation, Tarski keeps using the definite article ('*the* concept of truth') and the term 'true sentence', rather than a term of the form 'true in L', because he in effect treats 'true sentence' (as well as other terms, such as 'sentence' and 'the sign of the universal quantifier', etc.) as context sensitive: the term 'true sentence' expresses one concept, the concept *true sentence*, whose extension varies depending on which language is the salient one in a given context: in the context of much of CTF its extension is the set of true sentences of Calish; in other contexts, its extension might be the set of true sentences of E_0, or of G_0.[24]

---

should be possible for there to be different semantic systems of the same language, so that one would naturally expect Carnap to talk in terms of 'true in *S* of *L*'. But as one reads on, it turns out that Carnap tends to individuate languages in terms of semantic systems, which is why he has only 'true in *S*'. Note that this is doubly different from Tarski: first, Tarski does not use terms with built-in parameters anyway; second, if he did, it would not be Carnap's 'true in *S*'—after all, a semantic system contains metalinguistic rules and is defined in semantic terms.

[23] Carnap gives a version of Convention T in the form: "A predicate pr_i in [a metalanguage] *M* is an *adequate* predicate (and its definition an adequate definition) for the concept of *truth* with respect to an object language $S =_{Df} \ldots$" (Carnap 1942: 27). Three pages earlier, Carnap has said that the rules of truth of a system define 'true in *S*'. Note how Carnap seems to assume that defining *truth* with respect to *S* amounts to defining 'true in *S*''. (Note also the uncertain status of '*S*': Does it indicate a language or a semantic system of a language?)

[24] The view that 'true' is context sensitive is suggested, albeit somewhat indirectly, by Parsons (1974); it is explicitly advanced and worked out in more detail by Burge (1979). But both authors focus specifically on

I claimed earlier, by way of motivating the alternative interpretation, that Tarski does not use terms of the form 'true in L'. There is one possible exception to this claim, namely the very title of §3 of CTF, "The Concept of True Sentence in the Language of the Calculus of Classes" (CTF: 186), which can be parsed in the 'true in L' way, i.e. as talking of the concept *truth in Calish*. However, it can also be understood along the lines of the alternative interpretation, i.e. as talking of the concept *true sentence* defined for the language Calish. In favor of the second reading one can point out that the title is constructed in analogy to the title of the whole article, "The Concept of Truth in Formalized Languages", which is not to be understood in the 'true in L' way, for according to the standard interpretation there is no such concept as the concept *truth in formalized languages*.

Let us look at a few passages from Tarski that seem to bear on the two competing interpretations. In the Introduction of CTF, Tarski talks of *the* meaning of the term 'true' and of *the* concept of truth; he also says this:

> The extension of the concept to be defined depends in an essential way on the particular language under consideration. The same expression can, in one language, be a true statement, in another a false one or a meaningless expression. There will be no question at all here of giving a single general definition of the term. The problem which interests us will be split into a series of separate problems each relating to a single language. (CTF: 153)[25]

On the standard interpretation, Tarski is speaking rather misleadingly at the beginning of this passage: there is no such thing as *the* concept to be defined; instead, there are different concepts with different extensions—when Tarski is talking about "separate problems", he is talking about constructing different definitions of different concepts. On the alternative interpretation, he is not talking misleadingly. There is such a thing as *the* concept to be defined, the concept *true sentence*—when he is talking about "separate problems", he is talking about constructing different definitions of the same concept but *for* different object-languages.[26]

Late in CTF, towards the end of §5, Tarski adds some remarks about cases where whole classes of object-languages, instead of one single object-language, are under consideration. He says:

---

the behavior of 'true' in liar reasoning. The non-standard interpretation of Tarski under consideration here would maintain that a form of contextualism about 'true' and *truth* is suggested throughout CTF by Tarski's persistent use of 'true sentence', without built-in parameter, and of 'the concept of truth', without parameter but with the definite article.

[25] The occurrence of the word 'statement' in this passage is a bit disconcerting, for Tarski's 'true' is supposed to apply to sentences. However, my Polish informant tells me that 'statement' is a contribution by the translator: the Polish original (Tarski 1933) has 'zdanie', which corresponds exactly to English 'sentence' and is the term Tarski always uses in connection with 'true'. (The choice of 'Aussage' for 'zdanie' in the German translation (Tarski 1935) is quite unfortunate.) Thanks to Dr. Arkadiusz Chrudzimski, University of Salzburg.

[26] A similar passage can be found in one of Tarski's later writings. Having announced that he will apply the term 'true' to sentences, he says: "Consequently, we must always relate the notion of truth, like that of a sentence, to a specific language; for it is obvious that the same expression which is a true sentence in one language can be false or meaningless in another" (Tarski 1944: sec. 2). Note how, on the standard interpretation, the reference to *the* notion of truth is misleading; not so on the alternative interpretation.

> As I have already emphasized in the Introduction, the concept of truth essentially depends, as regards to both extension and content, upon the language to which it is applied. We can only meaningfully say of an expression that it is true or not if we treat this expression as a part of a concrete language. As soon as the discussion concerns more than one language the expression 'true sentence' ceases to be unambiguous. If we are to avoid this ambiguity we must replace it by the relative term 'a true sentence with respect to the given language'. (CTF: 263)

Note again Tarski's use of the definite article: he talks of *the* concept of truth as depending on the language to which it is applied. He also says the language dependence in question pertains to both the extension and *the content* of the concept of truth. Taken literally, this implies a distinction between the content of the concept and the concept itself.[27] If content can be equated with *intension* (cf. Tarski 1944: sec. 3), the passage can be taken to indicate that Tarski is committed to the transition 'different extension → different intension' but not to the transition 'different intension → different concept', hence not to the transition 'different extension → different concept'. These aspects of the passage suggest the alternative rather than the standard interpretation.

On the other hand, Tarski also says in the passage that the term 'true sentence' becomes *ambiguous* when more than one object-language is under consideration, and he refers to the disambiguated term as a *relative* term. These remarks might suggest the standard 'true in L' interpretation. However, things are not very clear-cut here. As to the remark that the term 'true sentence' becomes *ambiguous* when more than one object-language is under consideration, it depends on what Tarski means by 'ambiguous'. If he means that different occurrences of the term expresses different concepts, then the remark points towards the standard interpretation. If he merely means that different occurrences of the term have different extensions (or intensions), then the remark is compatible with the non-standard interpretation: if 'true sentence' is contextual, then different occurrences of the term can have different extensions (and even different intensions) while expressing one and the same concept.

As to the remark about the disambiguated term being *relative*, the continuation of the passage shows that Tarski is thinking there of constructing a single metalanguage common to the object-languages under consideration. He seems to be saying that, with such a metalanguage in hand and provided the object-languages are well-behaved ones, we should be able to define a genuine relational term, '*x* is true in language *y*', albeit one whose range of application, i.e. the range of the variable '*y*', will be restricted to the object-languages under consideration (cf. CTF: 263-4). This does not really fit well with the 'true in L' interpretation on which 'true in L' is not a relational term at all but merely a stand-in for various one-place predicates. One might also note that Tarski's remark about the disambiguated relative term is programmatic. He goes on to point out that "quite new complications might arise" when attempting to construct a definition of such a relative term, and he mentions specifically complications connected with "the necessity of defining the word 'language'" (CTF: 263-4). Again, this does not fit smoothly with the

[27] The actual phrase "the content of the concept of truth" shows up only in one place in CTF, namely in the paragraph preceding Convention T. In Tarski's 1944, sec. 3, we find the phrase "the meaning (or the intension) of the concept of truth".

'true in L' interpretation on which one would expect Tarski to mention difficulties connected with the notion of a language much earlier in his article and in a more prominent place.[28]

In sum, it seems to me quite difficult to tell whether Tarski's own intentions are better represented along the lines of the standard interpretation or along the lines of the alternative interpretation. Judging from how Tarski typically expresses himself, there is quite a bit to be said for the latter—though the evidence doesn't seem to be conclusive. The issue concerns the question of *concept individuation*. Tarski provides us with guidelines for constructing different definitions: Does he think these definitions will define one concept, *truth* or better *true sentence*, but define it *for* different object-languages, or does he think the different definitions will define different concepts of the form '*truth in L*'? As I said, it is difficult to tell. Tarski's later paper, "The Semantic Conception of Truth", contains an endnote suggesting that this question may, in the end, have no answer:

> The words "notion" or "concept" are used in this paper with all of the vagueness and ambiguity with which they occur in philosophical literature. Thus, sometimes they refer simply to a term, sometimes to what is meant by a term, and in other cases to what is denoted by a term. Sometimes it is irrelevant which of these interpretations is meant; and in certain cases perhaps none of them applies adequately. While on principle I share the tendency to avoid these words in any exact discussion, I did not consider it necessary to do so in this informal presentation. (Tarski 1944; endnote 4)

With respect to the last remark, we may observe that, while Tarski's CTF surely aims to be an "exact discussion", it does not exhibit much of a tendency to avoid the vague and ambiguous word 'concept'.

4. A Convention?

Tarski's Convention T is commonly described and treated as a condition of adequacy for a definition of truth. But Tarski labels it a *convention* and phrases it accordingly, using the words "will be called an adequate definition of truth". So, taken literally and seriously, Convention T does not state a condition under which something actually *is* an adequate definition of truth, it merely states a condition under which something *will be called* an adequate condition of truth.

Somewhat curiously, Tarski's does not comment at all in CTF on why he gives his condition the form of a convention. He does not use conventionalist language elsewhere in CTF (except for the sentence that introduces the convention, where he refers to it as a "postulate"); and his practice seems to belie to some extent his labeling and wording of Convention T: a reader of CTF will come away with the overall impression

[28] In the second half of §3 of CTF, after he has constructed his definition of 'true sentence' for Calish and has proved various theorems involving this term, Tarski refers to the defined concept as "the absolute concept of truth" and proceeds to define and discuss "another concept of a relative character", namely "the concept of *correct or true sentence in an individual domain a*", which applies, roughly speaking, to sentences that would be true in the ordinary sense if the quantifiers of the object-language under consideration were restricted to range over individuals of domain *a* (CTF: 199). Note that the relativization involved in '*x* is true in domain *a*', where the variable '*a*' ranges over sets of individuals to be associated with the quantifiers of Calish, is along a quite different dimension than the relativization involved in '*x* is true in language *y*', where the variable '*y*' ranges over an array of object-languages.

that Tarski intended to do rather more than merely recommend a convention. Since, moreover, he does not use the label 'Convention T' again in his later writings and calls the version he gives in "The Semantic Conception of Truth" a *criterion* for material adequacy (cf. Tarski 1944: sec. 4), one might even think that he was not serious when presenting it as a convention in CTF. But this would be rash. Although he drops the label, he keeps the conventionalist wording, using phrases such as "we shall say", "we will consider…as adequate", and "we stipulate", when presenting versions of his condition after CTF.[29]

Why does Tarski present his condition as a mere convention? One could try to understand this element of conventionalism as a reflection of his view about the indefinability of our concept of truth. The idea would be roughly this. According to Tarski, there cannot be a definition of a term '*Tr*' that actually *is* as a formally correct and materially adequate definition of *the* concept *true sentence*, that is, of *our* concept *true sentence*—that concept is not definable. Since the best we can hope for is various definitions of various *Ersatz*-concepts, it would be pointless to give a condition under which a definition actually *is* an adequate definition of truth. What Tarski does instead is to present a condition for *calling* a definition that assigns some other concept to '*Tr*' an "adequate definition of truth" (and to intimate further conditions for calling further definitions that assign still other concepts to '*Tr*' "adequate definitions of truth"). In effect, Convention T constitutes the recommendation to refer to some concept with our familiar term 'truth', even though the concept is not the/our concept *true sentence*, on the grounds that the concept is sufficiently similar to our concept to function as an Ersatz-concept.

This sort of account of the conventionalist aspect of Convention T would fit the standard interpretation mentioned earlier, on which Tarski shows us how to construct definitions of various Ersatz-concepts of the form '*truth in L'*, none of which is the/our concept *true sentence*. The account does not fit the alternative interpretation, on which Tarski *does* show us how to construct various definitions of the/our concept *true sentence*, albeit definitions that define that concept *for* certain object-languages—the concept being indefinable for various other object-languages, e.g. languages as rich as our ordinary languages. This might be counted as a point in favor of the standard interpretation. However, Tarski's own remarks from his later writings do not indicate any connection between the conventionalist aspect of Convention T and his views on the indefinability of truth.

In Tarski's late paper "Truth and Proof", we find some remarks concerning the goal and the logical status of an explanation of the meaning of a term. Tarski observes there that at times such an explanation "may be intended as an account of the actual use of the term involved", while at other times such an explanation "may be of a normative nature, that is, it may be offered as a suggestion that the term be used in a definite way" (Tarski 1969: 102). He then says that the explanation of the meaning of 'true' he wants to

[29] Compare Tarski's 1936, p. 404; 1944, secs. 4 and 9; 1969, pp. 106 and 114. When Tarski raises an issue of material adequacy in one of his earlier papers, "On Definable Sets of Real Numbers", he treats it as a factual, not as a conventional issue: "Now the question arises whether *the definitions just constructed…are also adequate materially*; in other words *do they in fact grasp the current meaning of the notion as it is known intuitively?*" (Tarski 1931: 128-9). But note that the definitions he is concerned with there are not definitions of truth.

give "is, to an extent, of mixed character"; and he continues: "What will be offered can be treated in principle as a suggestion for a definite way of using the term "true," but the offering will be accompanied by the belief that it is in agreement with the prevailing usage of the term in everyday language" (Tarski 1969: 102).

So far, one could maybe still see these remarks as being motivated along the lines of the account sketched above. But, as one reads on, it turns out that something else is on Tarski's mind. He reminds us that there are different *conceptions* of truth—the classical Aristotelian conception, the pragmatist conception, and the coherentist conception—and announces that he aims for the first one: "We shall attempt to obtain here a more precise explanation of the classical conception of truth" (Tarski 1969: 103).[30] As Tarski presents things, he conveys the impression that it is this *choice*—the choice to make precise the classical rather than some other conception of truth—that motivates him to put his condition into a conventionalist format. This is foreshadowed slightly at one point in CTF, where he says (albeit early on and long before he lays down Convention T) that he will be "concerned exclusively with grasping the intentions which are contained in the so-called *classical* conception of truth ('true—corresponding with reality') in contrast, for example, with the *utilitarian* conception (true—in a certain respect useful)" (CTF: 153). It comes out a bit more clearly—though it is not made explicit—in "The Semantic Conception of Truth", where he first mentions different conceptions of truth, then says he wants his definition "to do justice to the intuitions which adhere to the *classical Aristotelian conception of truth*" (Tarski 1944: sec. 3), and then proceeds to formulate a preliminary version of his condition/convention with reference to the classical conception: "…if we base ourselves on the classical conception of truth, we shall say…" (Tarski 1944: sec. 4). Moreover, in a later part of this paper, he briefly discusses the question whether the conception of truth he focuses on, the *semantic* conception, which is supposed to be a modernized form of the classical conception, is the "right" one. He professes "not to understand what is at stake in such disputes", urges us "to reconcile ourselves with the fact that we are confronted, not with one concept, but with several different concepts which are denoted by one word", and maintains that "the only rational approach to such problems" is to "try to make these concepts as clear as possible" (Tarski 1944: sec. 14). Note the indication that, as far as his investigation is concerned, he has *chosen* to make precise the concept characterized by the classical conception of truth.

So, judging from indications present in Tarski's own works, the conventionalist aspect of Convention T seems intended to reflect that a choice has been made by Tarski, that he has chosen to make precise the classical conception of truth rather than some other conception. If this is indeed the case, then the conventionalist aspect of Convention T is motivated by considerations that appear to be quite neutral between the standard and the alternative interpretation. At least as far as I can see, this motivation does not seem to favor either one of the two interpretations. The difference between them can be rephrased: Is it Tarski's intention to show us how to define different concepts of the form '*truth in L*', each of which is an Ersatz for the concept intended by the classical/semantic conception of truth, or is it his intention to show us how to define, albeit *for* different languages, the one concept, *truth* or rather *true sentence*, intended by the

[30] Tarski seems to distinguish implicitly between the *concept* of truth and a *conception* of truth. When he talks about a conception he has in mind something taking propositional form—a rough principle or definition that aims to tell us what truth is.

classical/semantic conception?

References

Burge, Taylor: 1979. Semantical Paradox. *The Journal of Philosophy* 76. Reprinted in R. L. Martin, ed., *Recent Essays on Truth and the Liar Paradox*, Oxford: Clarendon Press 1984: 83-117.

Carnap, Rudolf: 1942. *Introduction to Semantics*. Cambridge, Mass.: Harvard University Press.

David, Marian: 1996. Analyticity, Carnap, Quine, and Truth. *Philosophical Perspectives, 10, Metaphysics*: 281-96.

Gupta, Anil and Belnap, Nuel: 1993. *The Revision Theory of Truth*. Cambridge, Mass.: The MIT Press.

Parsons, Charles: 1974. The Liar Paradox. *Journal of Philosophical Logic* 3. Reprinted in R. L. Martin, ed., *Recent Essays on Truth and the Liar Paradox*, Oxford: Clarendon Press 1984: 9-45.

Patterson, Douglas, E.: 2006. Tarski on the Necessity Reading of Convention T. *Synthese* 151: 1-32.

Quine, W. Van: 1951. Two Dogmas of Empiricism. Reprinted in *From a Logical Point of View*, 2nd ed., revised, Cambridge, Mass.: Harvard University Press 1980: 20-46.

Tarski, Alfred: 1986 = CTF. The Concept of Truth in Formalized Languages. In *Logic, Semantics, Metamathematics*, translated by J. H. Woodger, 2nd ed., edited by J. Corcoran, Indianapolis: Hackett Publishing Company: 152-278. (First edition by Oxford University Press, 1956.) Revised translation of Tarski 1935.

—— 1969. Truth and Proof. *Scientific American* 220, June: 63-77. Reprinted in R. I. G. Hughes, ed., *A Philosophical Companion to First-Order Logic*, Indianapolis: Hackett 1993: 101-25. Page references are to the reprint.

—— 1946. *Introduction to Logic*. 2nd ed. New York: Oxford University Press. (First edition 1941.)

—— 1944. The Semantic Conception of Truth. *Philosophy and Phenomenological Research* 4: 341-75.

—— 1936. The Establishment of Scientific Semantics. Reprinted in *Logic, Semantics, Metamathematics*, 2nd ed., Indianapolis: Hackett Publishing Company 1986: 401-8. (Polish original in *Przegląd Filozoficzny*, vol. 39: 50-7.)

—— 1935. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* I, Lemberg: 261-405. Expanded translation of Tarski 1933.

—— 1933. *Pojęcie prawdy w językach nauk dedukcyjnych* (On the concept of truth in languages of deductive sciences): Warsaw.

—— 1931. On Definable Sets of Real Numbers. Reprinted in *Logic, Semantics, Metamathematics*, 2nd ed., Indianapolis: Hackett Publishing Company 1986: 110-42.