Forthcoming in: *New Essays on Tarski and Philosophy*. (Douglas Patterson, Ed.) Oxford University Press.

(Contributors: Jody Azzouni, Arianna Betti, Marian David, John Etchemendy, Solomon Feferman, Greg Frost-Arnold, Mario Gomez-Torrente, Wilfrid Hodges, Paolo Mancosu, Roman Murawski, Panu Raatikainen, Gila Sher, Peter Simons and Jan Wolenski.)

PANU RAATIKAINEN

TRUTH, MEANING, AND TRANSLATION

Philosopher's judgements on the philosophical value of Tarski's contributions to the theory of truth have varied. For example Karl Popper, Rudolf Carnap, and Donald Davidson have, in their different ways, celebrated Tarski's achievements and have been enthusiastic about their philosophical relevance. Hilary Putnam, on the other hand, pronounces that "[a]s a philosophical account of truth, Tarski's theory fails as badly as it is possible for an account to fail." Putnam has several alleged reasons for his dissatisfaction,[1] but one of them, the one I call the modal objection (cf. Raatikainen 2003), has been particularly influential. In fact, very similar objections have been presented over and over again in the literature. Already in 1954, Arthur Pap had criticized Tarski's account with a similar argument (Pap 1954). Moreover, both Scott Soames (1984) and John Etchemendy (1988) use, with an explicit reference to Putnam, similar modal arguments in relation to Tarski. Richard Heck (1997), too, shows some sympathy for such considerations. Simon Blackburn (1984, Ch. 8) has put forward a related argument against Tarski. Recently, Marian David has criticized Tarski's truth definition with an analogous argument as well (David 2004, p. 389-390).[2]

This line of argument is thus apparently one of the most influential critiques of Tarski. It is certainly worthy of serious attention. Nevertheless, I shall argue that, given closer scrutiny, it does not present such an acute problem for the Tarskian approach to truth as many philosophers think. But I also believe that it is important to understand clearly why this is so. Moreover, I think that a careful consideration of the issue illuminates certain important but somewhat neglected aspects of the Tarskian approach.

1. THE MODAL OBJECTION

The basic idea of the modal objection is simple enough: Instances of T-schema such as

'Snow is white' is true if and only if snow is white

are, in Tarski's approach, logical consequences of the truth definition and thus necessarily true; but certainly it would have been possible, so the argument goes, that 'snow' denoted, say, grass, in which case it would have been false that 'snow is white' is true if and only if snow is white. In other words, surely the sentence '"Snow is white" is true if and only if snow is white' is a contingent, empirical claim whose truth value depends on what the expressions of the object language mean, not a necessary truth, as

Tarski's approach entails. So, it is concluded, there must be something deeply wrong with Tarski's approach. In what follows, I shall focus mainly on Putnam's version of the modal objection, for Putnam has developed the argument in certain respects further than others (see Section 4), and considering those further developments allows one to clarify some interesting additional issues. I think that to the extent that Putnam's arguments can be rebutted, this should suffice also for the other variants of the modal objection.

In his much-cited "Comparison of Something with Something Else" (Putnam 1985: see also Putnam 1988), Putnam begins his modal objection by considering the following instance of T-schema:

(1) (For any sentence X) If X is spelled S-N-O-W-SPACE-I-S-SPACE- W-H-I-T-E, then X is true in L if and only if snow is white.

Putnam then presents his objection: "Since [(1)][3] is a theorem of logic in meta-L (if we accept the definition – given by Tarski – of 'true-in-L'), since no axioms are needed for the proof of [(1)] except axioms of logic and axioms about spelling, [(1)] holds in all possible worlds.[4] In particular, since no assumptions about the use of expressions of L are used in the proof of [(1)], [(1)] holds true in worlds in which the sentence 'Snow is white' does not mean that snow is white." (Putnam 1985, p. 333). Putnam concludes: "The property to which Tarski gives the name 'True-in-L' is a property that the sentence 'Snow is white' has in every possible world in which snow is white, *including worlds in which what it means is that snow is green* … A property that the sentence 'Snow is white' would have (as long as snow is white) no matter how we might use or understand that sentence isn't even doubtfully or dubiously 'close' to the property of truth. It just isn't truth at all." (Putnam 1985, p. 333).

John Etchemendy (1988), although reluctant to accept Putnam's most colorful conclusions, says that they are based on a "sound observation": "Tarski's definition does not provide an analysis of one important notion of truth" (p. 60, fn 8). More generally, he concludes that "the theory of truth that results from a Tarskian definition of truth … cannot possibly illuminate the semantic properties of object language" (Etchemendy 1988, p. 56). The reason Etchemendy gives for these claims is just the modal objection.[5]

2. Convention T and Translation

In order to evaluate the modal objection properly, one needs to take a closer look at Tarski's criterion of material adequacy, that is, his famous *Convention T*. It may be formulated as follows (cf. Tarski 1935, p. 187-8):

> A formally correct definition will be called *an adequate definition of truth* if it has the following consequences:
> (a) all sentences
> (T) *X* is true if and only if *p*,
> where '*X*' is a structural-descriptive name of a sentence *S* of the object language L and '*p*' is a *translation* of that sentence *S* into the metalanguage ML.

(b) for all *X*, if *X* is true, then *X* is a sentence of the object language L.

The reference to translation in (a) is important, although is often ignored, presumably because the more popular texts by Tarski (e.g. Tarski 1944) deal only with the case where the object language is assumed to be a (proper) part of the metalanguage (as in the standard example '"Snow is white" is true if and only if snow is white'); but it is essential to recognize that in this case it is tacitly assumed that the translation from the object language to the metalanguage is the trivial 'homophonic' one. If, on the other hand, one changes the interpretation of the symbols of the object language (with the result, say, that 'white' denotes green), the translation is no longer homophonic and must be made explicit. In his seminal paper on the concept of truth, Tarski was quite clear about these matters:

> We take the scheme [*x* is a true sentence if and only if *p*] and replace the symbol '*x*' in it with the name of the given sentence, and '*p*' by its translation into the metalanguage. (Tarski 1935, p. 187)

Instances of the schema (T) are nowadays often called T-sentences. As far as I know, this talk of T-sentences originated with Davidson (1973a, b). Note then that *if*, in a sentence of the form '*X* is true if and only if *p*', either:

(i) '*X*' is not a structural-descriptive name of *S*; or
(ii) '*p*' is not a translation of *S*,

then the equivalence '*X* is true if and only if *p*' does *not* count as an instance of T-schema, in other words, it is *not* a T-sentence.[6] Consequently, if one changes the interpretation of the symbols of the object language L, a former T-sentence may not be an instance of T-schema any more. That is, properly understood, Convention T necessarily requires that the relations between the object language L and the metalanguage ML be fixed (and remain constant). Let us try to see in more detail why this is so.

3. The Object Language as an Interpreted Language

As Tarski always insisted, truth can be only defined (because of paradoxes and Tarski's undefinability theorem) for a particular formalized language at a time. Moreover, for Tarski the 'formalized languages'[7] whose truth is under consideration always had to be interpreted languages,[8] as he repeatedly emphasized:

> It remains perhaps to add that we are not interested here in 'formal' languages and sciences in one special sense of the word 'formal', namely sciences to the signs and expressions of which no meaning is attached. For such sciences the problem here discussed has no relevance, it is not even meaningful. We shall always ascribe quite concrete and, for us, intelligible meanings to the signs which occur in the languages we shall consider. (Tarski 1935, p. 166-7)

Furthermore, this was, for Tarski, not just an accidental philosophical opinion;[9] rather, it was an essential part of Tarski's whole approach to truth that the meanings of the object language must be fixed. Only so could a truth definition (applied to sentences) make any sense at all:

> For several reasons it appears most convenient to apply the term 'true' to sentences, and we shall follow this course.[footnote omitted.]
>
> 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, p. 342)

> We shall also have to specify the language whose sentences we are concerned with; this is necessary if only for the reason that a string of sounds or signs, which is a true or a false sentence but at any rate meaningful sentence in one language, may be a meaningless expression in another. (Tarski 1969, p. 64)

> … the concept of truth essentially depends, as regards 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 sentences with respect to the given language'. (Tarski 1935, p. 263)

Therefore, it is necessary in Tarski's setting to focus on an interpreted language with constant meanings. If one varies the interpretation of the symbols of the object language L, the language changes to a different language L′; and (because one can define a truth predicate only for a particular language – an interpreted language – at a time) a former truth definition (true-in-L) is not a truth definition for this latter language L′; a former T-sentence does not count any more as a T-sentence (because T-sentences are defined only relative to a particular truth definition), and wholly different sentences become instances of T-schema – e.g., assuming that 'white' denoted (in-L′) green, one should now have 'The sentence "Snow is white" is true-in-L′ if and only if snow is green', etc.

All this is in stark contrast to the way formal languages are viewed in mature model theory, even though Tarski also importantly influenced the development of the latter. That is, in model theory, a language L is a completely uninterpreted and syntactic formal language. An L-structure *W* is defined as a pair (*D*, *I*), consisting of the domain *D* and the interpretation function *I*. The latter maps the non-logical symbols of *L* to elements of *D* (that is, the function *I* maps individual constants to elements of *D*, predicates to subsets of *D*, etc.).[10] In changing the structure, one varies the interpretation, but the language L remains the same.

Let us note in passing that the interpretation function *I* establishes a link between the object language and a domain of extra-linguistic objects, and hence is a semantical concept in Tarski's sense (see also below); Hence, it would be somewhat problematic to presuppose it in the Tarskian definition of truth,[11] which should not according to Tarski presuppose any semantical notions; the meanings of the object language must thus get fixed in some other way. Accordingly, it is important not to conflate Tarski's philosophical project of defining truth *simpliciter*, and the model-theoretic notion of truth-in-a-model defined in the above setting; their different understanding of what a language consists of is particularly relevant. However, all too often these are not clearly distinguished, and many misunderstandings derive from this. In particular, I suspect that such a conflation partly explains the popularity and attractiveness of the modal objection.

To recap, Tarski's approach to defining truth proceeds in certain order: First, an interpreted language equipped already with its meanings is chosen as the object language. Second, one presents a definition of the truth predicate for this particular interpreted language. The truth predicate defined is relative to this language and its interpretation. Finally, one shows that the definition is materially adequate by deriving T-sentences, which are doubly relative to the interpretation of the object language. *As an expression of German* (understood as an interpreted language), "weiss" necessarily means (means-in-German) what it does, namely white, and the same holds for all other expressions. As we have seen, if "weiss" denoted green, or "schnee" denoted grass, for example, the language would not be German any more. The identity of a language, in Tarski's setting, essentially depends on meanings of its expressions. Consequently, the equivalence '"Schnee ist weiss" is true-in-German if and only if snow is white' is, and should be, necessary, for the truth predicate is tied to the particular interpreted language. (cf. Milne 1997).

Let us now reconsider the modal objection. It is certainly true that expressions can change their meaning, and that the language could have so evolved that, for example, 'white' would denote green. However, from the Tarskian point of view, that language would no longer be English or, in short, L (as an interpreted language supplied with its meanings) – even if it were syntactically identical with L. Call this latter language L′. Even in such a possible world, it would nevertheless be true that 'white' denotes-in-L white, and that 'Snow is white' is true-in-L, if and only if, snow is white. It would only be the case that 'white' denotes-in-L′ green, and that 'Snow is white' is true-in-L′, if and only if snow is green. In other words, '"Snow is white" is true-in-L, if and only if, snow is white' is indeed true in every possible world and thus necessary.

In sum, Tarski's definition of truth does, pace Putnam, depend in a sense also on the meaning and not only on the spelling. Namely, meaning is built into the Tarskian approach via interpretation of the object language. So it seems that Putnam's modal objection can be effectively rebutted by pointing out that there is an illegitimate change of object language in the midst of the argument. Many of the critical replies to Putnam have indeed made this point (see e.g. Garcia-Carpintero 1996, Fernandez Moreno 1992, 1997, Niiniluoto 1994, Halbach 2001, Woleński 2001), and as far as it goes, this reply is, I think, on the correct lines.

4. The Identification of the Object Language

The whole issue is not, however, that easy to bypass, for Putnam is in fact aware of this 'language change reply' – as it might be called – and he has a further objection to this line of reply – an objection of which most of his critics seem to be ignorant. In *Representation and Reality* (Putnam 1988), Putnam reports how he raised the modal objection in a conversation with Carnap in the early 1950s: he complained that it isn't a logical truth that the (German) word 'Schnee' refers to the substance snow, nor is it a logical truth that the sentence 'Schnee ist weiss' is true in German if and only if snow is white. Carnap's reply was, Putnam recalls, that everything depends on the way the name of the language – 'German' or whatever – is defined. "[I]n philosophy, Carnap urged, we should treat languages as abstract objects, and they should be identified (their names should be defined) by their semantical rules. When 'German' is defined as 'the language with such and such semantical rules', it is logically necessary that the truth condition for the sentence 'Schnee ist weiss' in German is that snow is white." (Putnam 1988, p. 63) Putnam tells us that he was not satisfied, but did not continue the argument: "What I thought but did not say was: And, pray, what semantical concepts will you use to state these 'semantical rules'? And how will those concepts be defined?" (Putnam 1988, p. 63) Putnam then goes on to argue in some detail that if one attempts thus to define a language, one needs to appeal to the concept of truth, and that this would make the language change reply circular (Putnam 1988, p. 63-65).

Carnap apparently thought that languages should be identified (their names should be defined) by their semantical rules, and it may be that this is begging the question.[12] But be that as it may, it is important to note that this is not Tarski's view. Tarski explicitly points out the difference here between his own approach and that of Carnap, according to which we regard "the specification of *conditions* under which sentences of a language are *true* as an essential part of the description of this language." (see Tarski 1944, p. 373, note 24; my emphasis). For Tarski, on the other hand, the interpreted object language is instead specified simply through its metalinguistic translation (see e.g. Tarski 1935, p. 170-71; cf. Fernandez Moreno 1992, 1997; Milne 1997, Feferman 2004). In accordance, Tarski described the metalanguage in the following ways:

> … the metalanguage contains both an individual name and a *translation* of every expression (and in particular of every sentence) of the language studied … (Tarski 1935, p. 172; my italics)

> … to every sentence of the language … 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*. (Tarski 1935, p. 187; my italics)

However, one could point out that Tarski's approach still assumes the notion of *meaning,* in the disguise of translation, or the sameness of meaning. Does this mean that, in the end of the day, Tarski fails to achieve his expressed aim, that is, to define truth without assuming any *semantical* concepts? It has been frequently suggested that this is indeed the case. However, this is not necessarily so. In order to see this, we need to recall what

Tarski meant by 'semantical'. Tarski's paradigmatic examples of semantical concepts were satisfaction, denotation, truth and definability (see Tarski 1935, p. 164, p. 193-4; 1936, p. 401). He explained his understanding of 'semantical concept' as follows:

> A characteristic feature of the semantical concepts is that they give expression to certain relations between the expressions of language and the objects about which these expressions speak, or that by means of such relations they characterize certain classes of expressions or other objects. (Tarski 1935, p. 252)

In contrast, I submit that it is possible to view translation, in this context, as a purely syntactic, effective mapping between two languages, without assuming any relations between either language and objects about which they speak. Translation, so viewed, is *not* a semantical concept in Tarski's sense, and does not presuppose truth or related notions (most importantly, satisfaction, by means of which the others can be defined).[13] Hence, it seems to be, after all, admissible for Tarski to presuppose such a notion of translation in his approach without begging the question (cf. Milne 1997; see also below).

To conclude, Putnam's contention that defining the interpretation of the object language necessarily requires the notion of truth for that language is unproven, and the modal objection can indeed be disarmed – without begging the question – by recognizing that in the Tarskian approach, the object language, as an interpreted language with the meanings of its terms and hence their translations into the metalanguage held fixed, must remain constant and is not to be varied.

5. A Closer Look at the Tarskian Truth Definition

Let us now look in more detail, with a particular example, on how exactly Tarski himself specifies the meanings of the object language expressions and gives a truth definition. That one can derive the instances of T-schema in the metatheory is due to careful stage-setting; specifically, as Field (1972) has emphasized, the Tarskian definitions of satisfaction and truth are based on prior definitions of denotation for individual constants and of application for predicate constants – in short, of *primitive denotation*.[14]

For example, let us assume that the object language L is a (semi-formal) fragment of German. A Tarskian definition of denotation for names then takes the form of a list:

$$\begin{aligned} &\textit{Denotes}_{\mathrm{L}}(x, y) \leftrightarrow \\ &\quad [(x = \ulcorner \textit{Frankreich} \urcorner \ \wedge \ y = \textit{France}) \ \vee \\ &\quad (x = \ulcorner \textit{Deutschland} \urcorner \ \wedge \ y = \textit{Germany}) \ \vee \\ &\qquad\qquad \vdots \\ &\qquad\qquad \vdots \\ &\quad (x = \ulcorner \textit{Köln} \urcorner \ \wedge \ y = \textit{Cologne})]. \end{aligned}$$

Note that the number of primitive proper names is finite; consequently, denotation for names can be explicitly defined in the metalanguage; i.e., $Denotes_L(x, y)$ can always be eliminated, and one can use the right-hand side of the equivalence, which is a formula of the unextended metalanguage (assumed to contain no semantical concepts), instead. An analogous definition can be given for predicates:

$Applies_L(x, y) \leftrightarrow$

$[(x = \ulcorner Stadt \urcorner \wedge City(y)) \vee$
$(x = \ulcorner Staat \urcorner \wedge State(y)) \vee$
$\vdots$
$\vdots$
$(x = \ulcorner Rund \urcorner \wedge Round(y))]$.

This is how Tarski in practice fixes the interpretation of the object language (more exactly, the interpretation of its primitive non-logical symbols). Surely such a list-like explicit definition, which makes primitive denotation eliminable, does not presuppose any semantical notions. This should remove any remaining doubts as to whether Tarski could nail down the meanings of expressions of the object language without leaning on semantical concepts. In fact, denotation and application could be subsumed under a more general notion of satisfaction (see Tarski 1935, p. 190, p. 194), but for expository purposes, it is useful to present them separately as above. (A list-like characterization of primitive denotation such as above may strike one as disappointingly shallow philosophically, and one may sympathize Field's (1972) demand for a more substantial account of denotation, but there is, logically speaking, nothing in principle wrong in Tarski's approach – it is not in any way question-begging or circular.)

The recursive definitions of satisfaction and truth are familiar (For simplicity, let us assume that the object language L contains, as logical constants, only ~ (negation), & (conjunction), and *E* (existential quantifier)). I shall use ~, &, *E*, for the object language symbols, and $\neg$, $\wedge$, $\exists$ for the respective metalanguage symbols (and I assume that the metalanguage has also $\vee$, $\rightarrow$, $\leftrightarrow$, and $\forall$). *A* and *B* are formulas of L, *n* is a name in L and *P* is a predicate in L. σ, τ are infinite sequences of objects, and $\sigma(j)$ ($\tau(j)$) is the j^{th} member of the sequence σ (of the sequence τ).

$Satisfies_L(\sigma, x) \leftrightarrow$

$[(x = \ulcorner P(n) \urcorner \wedge (\exists y)(Denotes_L(\ulcorner n \urcorner, y) \wedge Applies_L(\ulcorner P \urcorner, y)) \vee$
$(x = \ulcorner P(v_j) \urcorner \wedge Applies_L(\ulcorner P \urcorner, \sigma(j))) \vee$
$(x = \ulcorner A \,\&\, B \urcorner \wedge Satisfies_L(\sigma, \ulcorner A \urcorner) \wedge Satisfies_L(\sigma, \ulcorner B \urcorner)) \vee$
$(x = \ulcorner \sim A \urcorner \wedge \neg\, Satisfies_L(\sigma, \ulcorner A \urcorner)) \vee$
$(x = \ulcorner (Ex_i)\, A \urcorner \wedge (\exists \tau)[(\forall j)(j \neq i \rightarrow \tau(j) = \sigma(j)) \wedge Satisfies_L(\tau, \ulcorner A \urcorner))]]$.

Note that this is not an explicit but a recursive definition, for $Satisfies_L$ occurs also in the right hand side of the equivalence. It is, however, possible to turn it to an explicit definition, with a help of a little bit of set theory.[15] The definition of truth is then simple:

$$True_{L}(x) \leftrightarrow [\, x \text{ is a closed formula } \wedge\ (\forall\sigma)(Satisfies_{L}(\sigma, x)).$$

All these definitions at place, one can then see that all the instances of T-schema, such as:

$$[True_{L}(\ulcorner Stadt(Köln) \urcorner) \leftrightarrow City(Cologne)]\ ,$$

can be derived in the metatheory.

6. Different Interpretations of Semantical Definitions

Now just to what extent such T-sentences are either true by definition and necessary, or contingent (the question at stake in the modal objection), is certainly parasitic to the modal status of what I shall call D-sentences and A-sentences. That is, by D-sentences, I mean sentences such as:

$$(\forall x)[Denotes_{L}(\ulcorner Mond \urcorner, x) \leftrightarrow x = the\ moon],$$

and by A-sentences, analogously, sentences such as

$$(\forall x)\,[Applies_{L}(\ulcorner Rund \urcorner, x) \leftrightarrow Round(x)].$$

Note that just like T-sentences, all D- and A-sentences are, in the Tarskian approach, provable theorems in the metatheory (given the definitions) and apparently necessarily true (assuming that the metatheory contains only arithmetical or set-theoretical axioms as its non-logical axioms; cf. note 4). The fundamental question concerns the modal status of such sentences; the modal objection could now be rephrased as the complaint that it is certainly a contingent empirical fact that e.g. 'Mond' denotes moon in German, and not a necessary truth as Tarski's approach seems to entail. The detour through T-sentences is really redundant and makes the issue unnecessarily complex and opaque.

Now it is true that such D- and A-sentences come out as "true by definition" in the approach that Tarski's takes to primitive denotation, and are provable in the metatheory, because *Denotes* and *Applies* can be explicitly defined. However, we have seen above that this is, after all, exactly how it should indeed be. The two-part definition of primitive denotation is constitutive for L as an interpreted language, and D- and A-sentences are immediate consequences of these definitions. Although it is obviously not necessary that 'Mond', as a mere string of symbols and viewed purely formally and syntactically, denotes moon, it is nevertheless the case that as a word of the interpreted language L, it necessarily denotes the moon.

One can look at the definition of primitive denotation in two different ways.[16] First, one may take the definition as purely *stipulative*, such that it defines the artificial language L as an abstract entity under consideration. From this perspective, there is

nothing external for the definition to be right or wrong about. However, one may alternatively be interested in an actual, concrete natural language, e.g. German, or rather a suitable formalizable fragment of such a language, and attempt to capture by a definition the pre-existing denotation relation[17] of that language in the metalanguage.[18] The definition aims to be *usage reporting*. From this perspective, one may well conclude in some case that definition or not, it has got the facts wrong. If the definition contained, for example, as its part the clause

$$(x = \ulcorner \textit{München} \urcorner \wedge y = \textit{Munster}),$$

one would have all the reasons to protest that it just isn't the case in German that the denotation of 'München' is Munster – 'München' denotes Munich – and to revise the definition. Surely, nothing in the formal definition itself dictates how to view it, but it is certainly possible to take the latter attitude towards the definition (cf. Davidson 1990).

At this point, it is illuminating to recall Carnap's distinction between pure and descriptive semantics (see Carnap 1942, p. 11-15). *Descriptive semantics* is concerned with historically given natural languages, such as German, and is based on empirical investigation. *Pure semantics*, on the other hand, is analysis of semantical systems with artificial languages which are stipulatively defined. It is entirely analytic and without factual content. "Here we lay down definitions for certain concepts, usually in the form of rules, and study the analytic consequences of these definitions. In choosing the rules we are entirely free" (Carnap 1942, p. 13). And we have seen that according to Carnap, in philosophy one must confine oneself to pure semantics. For Carnap, pure and descriptive semantics seem to be largely independent projects.

Tarski made an analogous distinction between *descriptive* and *theoretical semantics*. (Tarski 1944, p. 365). By "descriptive semantics", he refers to the totality of investigations on semantic relations which occur in a natural language. Apparently by "theoretical semantics" Tarski means kind of study he is himself pursuing. Fernandez Moreno (1997) suggests that theoretical semantics as undertood by Tarski corresponds to pure semantics in the sense of Carnap. However, I find this slightly problematic, or at least misleading. Carnap apparently viewed (in pure semantics) the definitions of semantical relations as purely stipulative, that is, thought that such definitions stipulatively define the language in question, and are analytically true of it. The language here is an artificial, formal language – an abstract object arbitrarily defined by the stipulations that govern its semantical relations.[19]

So what about Tarski? It is true that Tarski constantly insisted that colloquial languages give rise to semantical paradoxes, and that truth can only consistently be defined for a formalized language. This has led many to assume that Tarski, just like Carnap, wanted to restrict his "theoretical semantics" exclusively to artificial formal languages – that it is not at all applicable to the real-life natural languages. The case with Tarski is, however, more complicated than that. We have seen above that formalized or not, the languages under consideration must, for Tarski, be 'concrete' and already interpreted, in other words, must come already equipped with 'concrete' meaning. This alone makes them quite different from the artificial formal languages as usually

understood. Tarski also thought that his semantical tools can be applied to restricted languages of various special sciences, say, of chemistry – as long as they do not contain semantical vocabulary.

Moreover, Tarski suggests that theoretical semantics is, after all, applicable to natural languages, if "only with certain approximation" (Tarski 1944, p. 365). That is, "the approximation consists in replacing a natural language (or a portion of it in which we are interested) by one whose structure is exactly specified, and which diverges from the given language 'as little as possible'". (Tarski 1944, 347). Similarly, Tarski writes, "if we translate into colloquial language any definition of a true sentence which has been constructed for some formalized language, we obtain a fragmentary definition of truth which embraces a wider or narrower category of sentences" (Tarski 1935, p. 165, fn 2). In fact, Tarski at one point emphasised that by "formalized languages", he "does not have in mind anything essentially opposed to natural languages"; and he continues: "On the contrary, the only formalized languages that seem to be of real interest are those which are fragments of natural languages (fragments provided with complete vocabularies and precise syntactical rules) or those which can at least be adequately translated into natural languages" (Tarski 1969, p. 68).

For Tarski, the main problem with colloquial languages was that they are semantically closed,[20] for it is this aspect of them that leads to antinomies. However, suitable (semantically open) fragments of natural language, with sufficiently specified grammar, were wholly acceptable for him as object languages for truth definitions. Tarski had only complaints against natural language taken in its entirety (cf. Woleński 1993). Tarski himself described his view of theoretical and descriptive semantics thus:

> The relation between theoretical and descriptive semantics is analogous to that between pure and applied mathematics, or perhaps to that between theoretical and empirical physics; the role of formalized languages in semantics can be roughly compared to that of isolated systems in physics. (Tarski 1944, p. 365)

As a consequence of all the above, it seems as if Tarski was, unlike Carnap, inclined to view the definitions of semantical relations as usage reporting. That is, Tarski was inclined to the think that his definitions ultimately attempt to capture the actual semantical relations to the world of (fragments of) existing natural languages, rather than being merely stipulative specifications of artificial formal languages. (Such languages, of course, can certainly still play a role in the usage-reporting project.)

7. On Truth Definitions and Truth Theories

If one slightly relaxes Tarski's requirement that we do not use *any* semantical concepts in the truth definition, instead of explicitly defining primitive denotation one can add $Denotes_L$ and $Applies_L$ as new primitive predicates to the metalanguage, and then extend the metatheory with all D- and A-sentences as axioms governing them. One can then either explicitly define satisfaction and truth (assuming some set theory) in terms of primitive denotation, *or* add $True_L(x)$ and $Satisfies_L(x, y)$ as additional primitive

predicates and turn the relevant definitions to axioms governing them; the exact details do not matter here, where we are interested primarily in primitive denotation. The result is a *theory* of primitive denotation and truth, not a definition, and the D- and A-sentences are axioms of the theory. From this perspective, it is easier (than with definitions) to look at the theory as attempting to describe the actual denotation relations of the real target language, here German, and one can consider the axioms as having, in a sense, *empirical* content – exactly what, in part, the advocates of the modal objection demand. The suggested axiom $(\forall x)[\mathit{Denotes}_{\mathrm{L}}(\ulcorner \mathit{München} \urcorner, x) \leftrightarrow x = \mathit{Munster}]$, for example, would then be, even if an axiom, just a false hypothesis which should be revised, if the object language is supposed be (a fragment of) German.

But isn't it essential to the Tarskian approach to be able to explicitly define all semantical concepts? Does not giving up this requirement reopen the threat of paradoxes? And did not Tarski himself expressly oppose axiomatic theories of truth? These are good and natural questions to ask. However, I think that they suggest an a bit oversimplified picture of Tarski's view. It is true that from the beginning, Tarski announces the intention explicitly to define truth without using any semantical concepts, and it is also true that he eventually succeeds in doing so. Moreover, the possibility of explicitly defining truth in a logico-mathematical metatheory with no semantical concepts certainly removes any worries of the possibility of antinomy. However, it seems to be a mistake to assume that for Tarski, the primary solution to paradoxes is and has to be the requirement of explicit definability of the semantical concepts (in contrast to what e.g. Soames (1984, 1999) and Etchemendy (1988) seem to suggest). Rather, for Tarski, the real source of paradoxes was the universality or the semantical closedness of a language, and accordingly, the principal solution is the clear distinction between the object language and the metalanguage. (cf. Heck 1997). Whether or not one is able, and prefers, to give explicit definitions is a further issue.[21]

Moreover, the consistency of the above axiomatic theory of primitive denotation is guaranteed, for it can be easily shown to be a conservative extension of the original, unextended metatheory; therefore, no paradoxes can possibly threaten it. Hence there is little reason to resist such a move, and it is indeed difficult to see any reason why Tarski would have doubted the consistency of this theory – given that the separation of the object language and the metalanguage is clearly respected. In fact, even the full axiomatic theory of truth and satisfaction is likewise a conservative extension of a suitable unproblematic metatheory.[22]

It must be granted that there are some passages in Tarski where he contrasts the axiomatic approach with the definitional approach, and makes some critical remarks on the former (see Tarski 1936, p. 405-406, cf. 1935, p. 257-8). One problem Tarski mentions is the question whether the axiomatic semantical theory is consistent. However, in the approach we have just discussed this is not at all a problem; the consistency of the theory is guaranteed. Furthermore, Tarski complains that an axiomatic theory would be "highly incomplete", and that "the choice of axioms always has a rather accidental character". But if we look closer what Tarski really says, it becomes apparent that he has in mind first and foremost the weak theory which consists in mere T-sentences, and possible *ad hoc* extensions of this theory (Tarski 1935, p. 257-8). The reasons he gives

do not thus seem to count against just any kind of axiomatic theory of truth. Consequently, it seems that Tarski would not have had any strong reasons to object to an axiomatic theory such as one described above, which is in effect just Tarski's definitions transformed to an axiomatic theory. It is really just a different way of looking at Tarski's truth definition, and does not bring with it anything essentially new. Moreover, arguably Tarski himself was well aware of the possibility of such a transformation of his truth definition into a theory (cf. Heck 1997).

In sum, it is possible, without betraying the spirit of Tarski's project, to transform the Tarskian truth definition to an axiomatic theory, which can be interpreted to have empirical content. However, this does not mean that the relevant axioms and theorems are contingent. They still are constitutive and essential for the language in question. Perhaps they could be taken as another example of necessary truths that are knowable only *a posteriori*.[23]

References

Belnap, N.D.: 1993, 'On Rigorous Definitions', *Philosophical Studies* 72, 115–46.

Blackburn, S.: 1984, *Spreading the Word,* Oxford University Press, Oxford.

Carnap, R.: 1942, *Introduction to Semantics*, Harvard University Press, Cambridge, MA.

David, Marian 2004, 'Theories of Truth', in *I. Niiniluoto, M. Sintonen and J. Woleński (eds.), The Handbook of Epistemology, Kluwer Academic Publishers*, 331-413.

Davidson, D.: 1973a, 'In Defence of Convention T', reprinted in D. Davidson, *Inquiries into Truth and Interpretation.* Oxford: Oxford University Press, 65-75.

Davidson, D.: 1973b, 'Radical Interpretation', reprinted in D. Davidson, *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 125-139.

Davidson, D.: 1990, 'The Structure and Content of Truth', *Journal of Philosophy* **87**, 279-328.
Etchemendy, J.: 1988, 'Tarski on Truth and Logical Consequence', *Journal of Symbolic Logic* 53, 51-79.

Etchemendy, J.: 1988, 'Tarski on Truth and Logical Consequence', *Journal of Symbolic Logic* 53, 51-79.

Feferman, S.: 2004, 'Tarski's Conceptual Analysis of Semantical Notions', in A. Benmakhlouf (ed.), *Sémantique et épistémologie*, Editions Le Fennec, Casablanca [distrib. J. Vrin, Paris], pp. 79-108.

Fernandez Moreno, L.: 1992, 'Putnam, Tarski, Carnap und die Warheit', *Gräzer philosophische Studien* 43, 33-44.

Fernandez Moreno, L.: 1997, 'Truth in Pure Semantics: A Reply to Putnam', *Sorites*, Issue #08, June 1997, 15-23.

Field, H.: 1972, 'Tarski's Theory of Truth', *Journal of Philosoph* **69**, 347-75.

Frost-Arnold, G.: 2004, 'Was Tarski's Theory of Truth Motivated by Physicalism?', *History and Philosophy of Logic* 25, 265-280.

Garcia-Carpintero, M.: 1996, 'What is a Tarskian Definition of Truth?', *Philosophical Studies* 82, 113-144.

Halbach, V.: 2001, 'How Innocent is Deflationism?', *Synthese* 126, 167-194.

Heck Jr, R.: 1997, 'Tarski, Truth and Semantics', *Philosophical Review* 106, 533-554.

Lepore, E. & K. Ludwig.: 2005, *Donald Davidson: Meaning, Truth, Language, and Reality*, Oxford University Press, Oxford, 2005

Milne, P.: 1997,'Tarski on Truth and Its Definition', in Childers, Kolár and Svoboda (eds.), *Logica '96: Proceedings of the 10th International Symposium*, Filosofia, Prague, 1997, 189-210.

Niiniluoto, I.: 1994, 'Defending Tarski against his Critics', in B. Twardowski and J. Woleński (eds.) *Sixty Years of Tarski's Definition of Truth*, Philed, Warsaw, 48-68.

Pap, A.: 1954, 'Propositions, Sentences, and the Semantic Definition of Truth', *Theoria* XX: 23–35.

Patterson, D.: 2006, 'Tarski, The Liar and Inconsistent Languages', *Monist*, Forthcoming.

Putnam, H.: 1960, 'Do True Assertions Correspond to Reality?', in H. Putnam, *Mind, Language and Reality, Philosophical Papers Vol. 2*, Cambridge University Press, Cambridge, 70-84.

Putnam, H.: 1983, 'On Truth', in L. Cauman et al. (eds.) *How Many Questions? Essays in Honour of Sidney Morgenbersser*, Hackett, Indianapolis, 35-56; page references to the reprint in H. Putnam, *Words and Life* (ed. J. Conant) Harvard University Press, Harvard, 1994, 315-329.

Putnam, H.: 1985, 'Comparison of Something with Something Else', *New Literary History*, 17, 61-79; page references to the reprint in H. Putnam, *Words and Life* (ed. J. Conant) Harvard University Press, Harvard, 1994, 330-350.

Putnam, H.: 1988, *Representation and Reality*, MIT Press, Cambridge.

Quine, W.V.: 1946, 'Concatenation as a Basis for Arithmetic', *Journal of Symbolic Logic*, vol. 11, 105--114.

Raatikainen, P.: 2003, 'More on Putnam and Tarski', *Synthese* 135, 37-47.

Soames, S.: 1984, 'What is a Theory of Truth', *Journal of Philosophy* 81, 411-29.

Soames, S.: 1999, *Understanding Truth*, Oxford University Press, Oxford.

Tarski, A.: 1935, 'The Concept of Truth in Formalized Languages', in A. Tarski: *Logic, Sematics, Metamathematics* (2nd edition) J. Corcoran ed., Hackett, Indianapolis, 1983, 152-278.

Tarski, A.: 1936, 'The Establishment of Scientific Semantics' in A. Tarski: *Logic, Sematics, Metamathematics* (2nd edition) J. Corcoran ed., Hackett, Indianapolis, 1983, 401-408.

Tarski, A.: 1944, 'The Semantic Conception of Truth and the Foundations of Semantics", *Philosophy and Phenomenological Research* 4, 341-376.

Tarski, A.: 1969, 'Truth and Proof', *Scientific American* 220 (June 1969), 63-77.

Woleński, J.: 1993, 'Tarski as a Philosopher', in F. Coniglione et al. (eds.) *Polish Scientific Philosophy: The Lvov-Warsaw School*, Rodopi, Amsterdam, 319-338.

Woleński, J.: 2001, 'In Defense of the Semantic Definition of Truth', *Synthese* 126, 67-90.

Notes

[1] See Putnam 1960, 1983, 1985, 1988. For criticism, see Raatikainen, 2003.

[2] As Halbach (2001) has pointed out, analogous arguments have been presented also by Lewy, Strawson, Church, and Quine, though not always directly as a criticism of Tarski.

[3] I have changed Putnam's numbering.

[4] Putnam's claim is exaggerated: in the standard cases, where there are infinitely many sentences, at least a weak subsystem of the second-order arithmetic such as ACA – and not just logic – is needed for the truth-definition and the derivation of T-sentences from it. However, as the great majority of philosophers apparently think that theorems of arithmetic also are necessary and *a priori*, and this is the crucial matter here, I shall not make more about this.
Thus let us assume that the *metatheory* does not contain any non-logical axioms except arithmetical axioms, or axioms of the theory of concatenation (or syntax), which amounts to the same (Quine (1946), for example, shows that elementary arithmetic and the elementary theory of concatenation are equivalent). The *metalanguage*, on the other hand, may and often must contain other sorts of non-logical expressions, such as 'green' 'moon', 'round', 'Earth' etc. in our examples; the point is that there are no non-logico-arithmetical axioms governing them. Under these assumptions, T-sentences are just definitional abbreviations of certain theorems of arithmetic, and thus, according to the standard view, indeed necessarily true and *a priori* knowable. Had the metatheory other sorts (e.g. contingent or empirical) of axioms, being a consequence of a definition would not make a theorem anything more than contingent.

[5] For Etchemendy's version of the modal argument, see Etchemendy 1988, p. 56-7, 60-61.

[6] There is much unclarity and confusion on this matter in the literature. Thus one often counts sentences such as '"Snow is white" is true iff the moon is made out of cheese' as T-sentences, and talks about false T-sentences. But such sentences simply are not T-sentences. I think one should call them e.g. alleged or apparent T-sentences, or T-like sentences (as Lepore & Ludvig (2005) do), in order to clearly distinguish them from the genuine T-sentences.

[7] One may also note that the title of the Polish original of 'The concept of truth in formalized languages' did not even speak about formalized languages, but translates in fact as 'The concept of truth in the languages of deductive sciences'.

[8] To be sure, certain characterizations of 'formalized languages' by Tarski are quite misleading and confusing, e.g., when he writes that formalized languages "can be roughly characterized as artificially constructed languages in which the sense of every expression is uniquely determined by its form" (Tarski 1935, p. 165-6).

[9] Apparently Tarski originally accepted this idea by accepting his teacher's Leśniewski's 'intuitionistic formalism', according to which all languages, including formal ones, are already interpreted (this was considered not to be an obstacle for their formalization). But Tarski still held this view much later (still in 1969), when he otherwise had distanced himself quite a lot from Leśniewski's philosophical ideas.

[10] Obviously, there are different ways to formulate these ideas, but in practice they are equivalent to the one presented here.

[11] Though, it is of course perfectly acceptable in its proper context, in model theory, whose aims are quite different.

[12] But see Fernandez Moreno 1997.

[13] It must be granted that that issue is not absolutely crystal clear. For example, in 1944 Tarski wrote: "Within theoretical semantics we can define and study some further notions, whose intuitive content is more involved and whose semantic origin is less obvious; we have in mind, for instance, the important notions of *consequence*, *synonymity*, and *meaning*." He adds (fn 20) that all those notions can be defined in terms of satisfaction; and refers to Carnap (1942) for the definition of synonymity. Doesn't this passage undermine my conclusion in the text? I am inclined to that that not. First, Tarski seems to be talking here about intralinguistic synonymity between two expressions of the object language L, and not about interlinguistic synonymity (translation) between L and ML. Second, Tarski only says that it is possible to define synonymity in terms of satisfaction; he does not state that it cannot be fixed in any other way. Third, he is here referring more to Carnap's work than to his own.

[14] For simplicity, I assume that L does not contain function symbols and that it only has monadic predicates.

[15] Or, alternatively, one can transform it to an axiomatic theory. This is relevant in what follows.

[16] For more about the difference between stipulative and usage reporting (or lexical) definitions, see e.g. Belnap 1993.

[17] More exactly, its restriction to the relevant fragment.

[18] Obviously, the way I have developed the truth definition above is already inclined towards this interpretation.

[19] Whether this is a completely fair interpretation of Carnap's views I am not sure – it may well be an oversimplified account (in any case, his later thoughts about explication suggest a more sophisticated view). However, this does not really matter; my aim here is to argue that Tarski did not hold the view I describe here – whether or not this is exactly the overall view of historical Carnap.

[20] Or, more accurately, that they purport to be semantically closed (see Patterson 2006).

[21] If, however, one takes seriously Tarski's once declared requirement of physicalistic acceptability of the semantic notions, the need of explicit definability may be more acute. However, I am inclined to think that physicalism was not really that essential to Tarski's project; the only context where he talks about it (Tarski 1936) was a popular presentation of his work for an audience with many logical positivists there. See also Frost-Arnold 2004.

[22] Not object theory. Assuming that the object language has at most the expressive power of the language of first-order arithmetic (of course, it may have nothing to do with arithmetic or mathematics), the weak subsystem of second order arithmetic ACA is sufficient for most purposes. The full axiomatic theory of truth over the language of first-order arithmetic, which allows induction scheme to be applied also to formulas which contain truth predicate, is equiconsistent with ACA.

[23] I am very grateful to Douglas Patterson for his valuable comments to an earlier version of this paper.