Forthcoming in SYNTHESE

PANU RAATIKAINEN

MORE ON PUTNAM AND TARSKI

ABSTRACT. Hilary Putnam's famous arguments criticizing Tarski's theory of truth are evaluated. It is argued that they do not succeed to undermine Tarski's approach. One of the arguments is based on the problematic idea of a false instance of T-schema. The other ignores various issues essential for Tarski's setting such as language-relativity of truth definition.

Hilary Putnam has repeatedly attacked Tarski's theory of truth. He famously concludes that Tarski's account fails as badly as it is possible for an account to fail. And when a philosopher and a logician as distinguished as Putnam is making such radical statements, they must be taken seriously.

Somewhat surprisingly, there have been relatively few critical responses to these bold claims in the literature (some lines of criticism by Putnam have remained almost totally unresponded). On the contrary, some of the key papers in the field (see e.g. Soames 1984, Etchemendy 1988, Heck 1997) refer to Putnam's conclusions more or less approvingly. Hence one may get the impression that Putnam has shown conclusively that there is something deeply wrong in Tarski's theory of truth. Nevertheless, it is my aim in this paper to argue that ingenious as these arguments by Putnam are, they miss their target. Although I feel no dogmatic need to defend Tarski, I think that Putnam's arguments do not in fact affect Tarski's theory – whatever its true limitations may be.

Putnam has two basic objections to Tarski's theory. First is based on the possibility of a metatheory that proves some false theorems but satisfies Convention T. I shall call this 'the unsoundness objection'. Another is based on certain modal considerations: T-sentences are necessary logical consequences of truth definition; but certainly it would have been possible that 'snow' denoted grass, in which case it would have been false that 'snow is white' is true if and only if snow is white.[1] I shall call this line of criticism 'the modal objection'.

There has been some critical discussion of the modal objection in the literature (Garcia-Carpintero 1996, Luis Fernandez Moreno 1992, 1997, Niiniluoto 1994, Halbach 2001, Wolenski 2001), some in my mind along the right lines, but none really conclusive (cf. Section 4).[2] The only reply to the unsoundness objection that I know is a recent paper by Jan Wolenski (2001). Wolenski provides rather involved considerations on ω-consistency, ω-completeness, non-standard models etc. But yet, Wolenski's treatment still leaves the issue open – he does not provide a conclusive refutation of Putnam's objection. And I don't think that he really manages to reveal what in my mind is the basic mistake in Putnam's reasoning. Consequently, I think that Putnam's arguments are worthy of one more reply.

1. The Unsoundness Objection

Let us first consider Putnam's unsoundness objection. It is directed against Tarski's material adequacy condition, i.e. his famous Convention T. Tarski required that all substitution instances of the following schema

(T) 'P' is L-true $\leftrightarrow$ P (where 'P' is a name of L-sentence P)

must be theorems of metatheory MT, if L-true is to be an adequate truth predicate for a language L. But according to Putnam, "this will not do": if MT is ω-inconsistent, then it may be the case that some instances of the above schema are theorems of MT although (in the metametalanguage) one can easily see that they are all not (in the intuitive sense) true. For, in general, "[t]he fact that all instances of a schema are provable in some language or other does not mean that all instances of that schema are true" (Putnam 1983, 319).

In another context (Putnam 1975, 73), Putnam formulated his objection more generally in terms of incorrectness (in contradistinction to more specific ω-inconsistency). That is, he invites us to suppose that MT is "incorrect"; presumably he means by this that MT proves some (in the intended sense) false theorems, or in other words, is unsound. He continues: "Then the biconditionals ['P' is L-true if and only if P] may all be theorems of [MT] even though some of them are false. In this case, the definition of 'true' would satisfy Tarski's Criterion of Adequacy even though it was extensionally incorrect!"

Putnam concludes that it is either false that the criterion of adequacy is materially sufficient or we must strengthen the criterion by requiring that all axioms of MT must be *true* – in which case it is false that the criterion of adequacy does not presuppose the notion of truth (Putnam 1975, 73).

For our present purposes, we need not worry too much about the difference between these two formulations: ω-inconsistency is a particular form of unsoundness, and the relevant issue in any case is Putnam's idea that MT may possibly prove some false instances of T-schema. How fatal is this for Tarski?

I think that Putnam's argument sounds cleverer than it actually is. For how could such a biconditional be false, however unsound MT is? That is, one may wonder what a false instance of T-schema would look like?[3]

Putnam assumes that an unsound MT proves a false case of T-schema, i.e. for some P,

'P' is true-in-L $\leftrightarrow$ P,

where either:

(a) ' "P" is true-in-L' is (in the intuitive sense) true but P is false;

or:

(b) ' "P" is true-in-L' is (in the intuitive sense) false but P is true.

(By the truth table of $\leftrightarrow$, these are the two possible cases in which the equivalence may be false.)

But in fact, neither case makes any sense, viz. neither case is actually possible. For, consider first (a) (we work within a (correct) metametatheory): if ' "P" is true-in-L' is (in the intuitive sense) true, then certainly P is true, not false; and analogously with (b). In sum, no T-sentence (i.e. an instance of T-schema) can be false. Thus, contrary to a widely held opinion,[4] I submit that T-sentences are in fact necessary.[5] (At this point, one may ask what if 'P' was not a name of P? Isn't this a counterexample for the necessity of T-sentences? No, it isn't, for the resulting sentence would not count as an instance of T-schema any more (see also below, Section 3).)

Note, however, that it is possible (indeed easy) to formulate a metatheory MT with a 'materially adequate' truth predicate Tr(x) such that it proves Tr('P') for some intuitively false P (in arithmetic, false in the standard model) – and is in this sense extensionally incorrect.[6] The simplest way to do this is to first add to a base theory B (say, Peano Arithmetic PA) a new predicate Tr(x) and all instances of Tr('P') $\leftrightarrow$ P, where P is a

sentence of the base language L(B); this extension is known to be consistent if the base theory B is, and is moreover conservative over the latter (that is, this extension does not prove any sentences of the base language L(B) which B itself doesn't prove). Second, one adds as a new axiom some false but unrefutable sentence F of the base language (e.g. the claim that the base theory is inconsistent, which is irrefutable by Gödel's second incompleteness theorem). The resulting metatheory MT trivially proves F, and by the relevant instance of T-schema, proves also Tr('F'). It may be that it is a case like this that Putnam had in mind, but that he confused this with the wholly different and quite puzzling idea of a false instance of T. But it is not the case that Tr('F') ↔ F is somehow false; it is true because both sides of the biconditional are false (in the intended intuitive sense).

Does an observation such as this contradict Tarski's statement that "the conditions for the material adequacy of the definition uniquely the extension of the term 'true'" (Tarski 1944, 353)? I don't think so. What Tarski apparently meant here is that given any two alternative truth-predicates Tr(x) and Tr*(x) both satisfying Convention T (i.e. for every L-sentence P, MT proves both Tr('P') ↔ P and Tr*('P') ↔ P), MT proves Tr('P') ↔ Tr*('P')), for every L-sentence P. However, it is important to note that this does *not* mean that the predicates are co-extensive in an arbitrary model, or that T-sentences implicitly define the set of true sentences (which can be refuted by combining Beth's theorem with Tarski's undefinability theorem; see e.g. Ketland 1999. Tarski was in fact aware of this from the beginning; see Tarski 1935, 258; cf. Ketland 1999).

Now certainly Tarski assumed that a theory used as a metatheory is a theory accepted ('as true') – that it is honestly believed in (cf. Coffa 1991, 296). But it is a theory which is used, not mentioned. And it would certainly be fantastic if in general, if one has a set of beliefs, a theory, it would require that one must possess concepts (or even definitions of the concepts) of the exact semantics of the language of the theory.

2. Other Circularity Objections

Putnam has also put forward different complaints of circularity against Tarski's theory of truth:

> However [the claim that the criterion of adequacy does not presuppose the notion of truth (or immediately related notions such as 'naming')] is false also for other (no less important)

> reasons: the Criterion of Adequacy presupposes the semantical notion of naming (['S'] must name S and not, e.g. the negation of S) and the semantic notion of a biconditional (a connective with a certain truth-table). (Putnam 1975, 73)

It is true that the general notion of 'naming' (or 'denotation'), which may be paraphrased as 'true of', is such a close relative of the notion of 'truth' – and indeed truth can be defined in terms of it (cf. McGee 1991, 32-33) – that if a definition of truth assumes it, the definition is viciously circular. So is Tarski's definition like this? No! One must note, against Putnam, that:

(i) the notion of naming does not occur in Tarski's definition of truth, but only in the Criterion of Adequacy, which is a test of a definition – and which is formulated only in the metametalanguage (MML).

(ii) The notion of naming used in the Criterion of Adequacy is very limited and simple; it is restricted to the primitive denotation of sentences by their canonical ('structural-descriptive') names. It does not require the general notion of denotation amounting to 'true of'. The former is a decidable relation, whereas the latter is not even arithmetical (assuming that the language contains elementary arithmetic; and because the language under consideration here is in fact the metalanguage, it certainly must be that rich). Analogously, the truth-table for biconditional does not go beyond decidable, and cannot possibly presuppose the general notion of truth.

3. The Modal Objection

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

(2) (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 (2) 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 (2) except axioms of logic and axioms about spelling, (2) holds in all possible worlds. In particular, since no assumptions about the use of expressions of L are used in the proof of (2), (2) holds true in worlds in which the sentence 'Snow is white' does not mean that snow is white." (Putnam 1985, 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, 333)

Note, by the way, that here Putnam admits that T-sentences are necessary, and thus cannot be false, and so contradicts the basic idea of his unsoundness objection; he now only claims that their being necessary makes Tarski's approach implausible.

In order to evaluate Putnam's argument properly, one needs to take a closer look at Tarski's criterion of material adequacy, that is, his **Convention T**. It may be formulated as follows (cf. Tarski 1935, 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 OL and 'p' is a translation of that sentence S into the metalanguage.
(b) *for all X, if X is true, then X is a sentence of OL.*

The issue of translation here is important although is often ignored, presumably because the more popular texts by Tarski 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 OL to ML is the trivial 'homophonic' one. If, on the other hand, one changes the interpretation of the symbols of OL (resulting, say, that

'w-h-i-t-e' denotes green), the translation is not any more homophonic and must be made explicit. In his seminal paper on the concept of truth, Tarski was quite explicit 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, 187)

Note then that *if* 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. Consequently, if one changes the interpretation of the symbols of object language, a former T-sentence is not an instance of T-schema any more. That is, Convention T necessarily requires that the relations between object-language and metalanguage be fixed (and remain constant). Let us try to see in more detail why this is so.

As Tarski always pressed, 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 whose truth is under consideration were, and had to be, always already interpreted languages, as he repeatedly insisted:

> I should like to emphasize that, when using the term 'formalized languages', I do not refer exclusively to linguistic systems that are formulated entirely in symbols, and I do not have in mind anything essentially opposed to natural languages. 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, 68)

> 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, 166-7)
>
> A formal system … for which we are unable to give a single interpretation, would, presumably, be of interest to nobody. (Tarski 1941, 129)

Furthermore, this was not just an accidental philosophical opinion from Tarski's side, but it is an essential part of Tarski's whole approach to truth that the meanings of the object language must be fixed. Only that way can 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, 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 sings, which is a true or a false sentences 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 sentences' 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.[7] If one changes the interpretation of the symbols of the object language, the language changes to a different language; a former truth definition is not a truth definition for this latter language, a former T-sentence does not count any more as a T-sentence, and wholly different sentences become instances of T-schema (in Putnam's

example, assuming that ‘white’ meant green, one should have ‘The sentence “Snow is white” is true if and only if snow is green’, etc.).

A truth-definition and T-sentences for it are thus necessarily relativized to a particular interpreted object language OL; if the interpretation is changed, the language is not the same any more. And once a language (with an interpretation) is fixed, T-sentences are indeed necessary, although it is obviously not necessary that the symbols of OL have to be interpreted in that way. T-sentences are necessarily true sentences of the metalanguage ML; the contingent talk of the meanings of the expressions of metalanguage ML (the denotation of the structural-descriptive names of sentences of OL) and of the relations (in particular, the translation) between the object language OL and the metalanguage ML (explicit in Convetion T as formulate above) belong only to the metametalanguage MML. Putnam’s argument confuses all these issues.

In sum, Tarski’s definition of truth does, pace Putnam, depend also on the meaning and not only on the spelling. Meaning is built into the Tarskian approach via interpretation and translation. For it is assumed that a translation of the object language OL into the metalanguage ML has been fixed. 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, and as far as it goes, this reply is, I think, on the correct lines.

4. Language and Truth

The whole issue is not, however, that easy, for Putnam is in fact aware of this ‘language change reply’, and he has a further objection to this line of reply. In *Representation and Reality* (Putnam 1988), Putnam reports how he raised the modal objection in a conversation with Carnap already 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’ of 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, 63)

Putnam tells 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, 63) To make a long story short: Putnam next argues that if one attempts thus to define a language, one needs to appeal to the concept of truth; and certainly this would make the language change reply circular (Putnam 1988, 63-65).

Interesting as these considerations are, I think that they do not in fact hit the target, if that is supposed to be Tarski. One reason is that whatever may be the problems with Carnap's position, it is odd to commit Tarski to these same views only because Carnap was inspired by Tarski's work. Putnam slides from Tarski to Carnap and back without making it clear whose position he is criticizing. Although in other contexts Putnam is attacking, presumably for these same reasons, expressly Tarski's theory, here he deals quite clearly with Carnap's specific views.

It may be that Carnap thought that languages should be identified (their names should be defined) by their semantical rules. But this was not Tarski's view; Tarski noted this difference between him and Carnap explicitly (see Tarski 1944, 373, note 24). For Tarski the object language, with an interpretation, is fixed simply by its translation to the metalanguage (cf. Fernandez Moreno 1992, Milne 1997). This is apparently an effective mapping, something which is a much more simple notion than truth, which is easily (if the object-language can express the basic arithmetical notions) even non-arithmetical (thus the notion of interpretation in question here cannot presuppose the concept of truth).[8] 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, 172)

> … 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, 187)

To conclude, Putnam's claim that defining (the interpretation of) the object language requires the notion of truth (for that language) is false, and the modal objection can be refuted by recognizing that object language, as an interpreted language with its meanings fixed, and its translation to the metalanguage must remain constant.[9]

REFERENCES

Boisvert, D.R.: 1999, 'The trouble with Harrison's "The trouble with Tarski"', *Philosophical Quarterly* 49, 376-383.

Coffa, J. A.: 1991, *The Semantic Tradition from Kant to Carnap: To the Vienna Station*, Cambridge University Press, Cambridge.

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

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.

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

Gupta, A.: 1978, 'Modal Logic and Truth', *Journal of Philosophical Logic* 7, 441-472.

Gupta, A. and N. Belnap: 1993, *The Revision Theory of Truth*, MIT Press, Cambridge.

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

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

Ketland, J.: 1999, 'Deflationism and Tarski's Paradise', *Mind* 108, 69-94.

Kirkham, R.: 1992, *Theories of Truth,* MIT Press, Cambridge.

Lewy, C.: 1947, 'Truth and Significance', *Analysis* 8, 24-27.

McGee, V.: 1991, *Truth, Vagueness, and Paradox,* Hackett, Indianapolis.

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.

Milne, P.: 1999, 'Tarski, Truth and Model Theory', *Proceedings of the Aristotelian Society* XCIX (1998-9), 141-167.

Niiniluoto, I.: 1994, 'Defending Tarski against his Critics', in B. Twardowski and J. Wolenski (eds.) *Sixty Years of Tarski's Definition of Truth*, Philed, Warsaw, 48-68.
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.
Soames, S.: 1984, 'What is a Theory of Truth', *Journal of Philosophy* 81, 411-29.
Soames, S.: 1995, 'T-Sentences', in W. Sinnott-Armstrong et al. (eds.) *Modality, Morality and Belief,* Cambrdige University Press, Cambridge.
Tarski, A.: 1935, 'The Concept of Truth in Formalized Languages', page references to the English translation in A. Tarski: *Logic, Sematics, Metamathematics* (2nd edition) J. Corcoran ed., Hackett, Indianapolis, 152-278.
Tarski, A.: 1941, *Introduction to Logic*, Oxford University Press, New York.
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.
Wang, H.: 1952, 'Truth Definitions and Consistency Proofs', *Transactions of American Mathematical Society* 73, 243-275.
Wolenski, J.: 2001, 'In Defense of the Semantic Definition of Truth', *Synthese* 126, 67-90.

NOTES

[1] Halbach (2001) points out that variants of this argument can be traced back to Lewy (1947) and that also Church, Quine, Strawson, Papp, Rescher etc. have presented similar objections.

[2] Moreover, some of the best replies are much less known and/or much less easily accessible; I have in mind Fernandez Moreno (1992), (1997), Milne (1997) (Milne (1999) also useful) and (Boisvert 1999), the two last mentioned make many of the relevant points although they are not written as replies to Putnam; Milne mentions Putnam's modal objection only by passing, and Boisvert is replying to Harrison, who had criticized Tarski in a way similar to Putnam's. Fernandez Moreno is the only critic of Putnam who takes in account Putnam (1988), but he focuses on defending Carnap.

[3] If the truth-predicate 'true-in-L' were allowed to occur in P in a T-sentence, then it would make sense to talk about false T-sentences. Let us write Tr(x) for the 'truth predicate'. One then diagonalizes with $\neg$Tr(x),

and gets a sentence X such that $X \leftrightarrow \neg Tr(\text{`X'})$, which in turn is equivalent to $\neg(Tr(\text{`X'}) \leftrightarrow X)$. Such a 'paradoxical' biconditional is certainly false, but note that it is not an instance of Tarski's Convention T, which requires that P is a sentence of the object language OL, whereas X here contains the truth-predicate Tr(x) and hence cannot possibly be a sentence of OL but only of ML.

[4] See e.g. Etchemendy (1988), Soames (1995), Kirkham (1992), Niiniluoto (1994).

[5] I thus agree with Gupta (1978), Belnap and Gupta (1993), Milne (1997) and Halbach (2001).

[6] The possibility of such unsound materially adequate MT has been known at least since Wang (1952).

[7] This important point is often confused because of the notion of truth-in-a-model, applicable to an uninterpreted formal language, which is a wholly different notion from the notion of (absolute) truth occurring in T-schema.

[8] There is an interesting but rarely (never?) noted complication here: Although Tarski's definition of truth (for a fixed OL) is given in the metalanguage ML, Tarski's setting requires one to rise to a metametalanguage MML at least at this one place. That is, T-sentences are sentences of ML, but already in order to know which sentences count as T-sentences, one must rise to MML (and, of course, when one then asks wither MT proves all instances of T-schema): viz. one must fix a translation between OL and ML, and this can be done only in MML. Is this a serious problem? It may appear that it is, for the talk about sentences of OL and ML having the same meaning would appear to already assume semantical concepts, and thus destroy Tarski's basic aim to define truth in purely non-semantical terms. However, on closer examination I think this turns out to be rather harmless, for a translation here, although given in MML, is – as noted – quite simple notion, a decidable relation between sentences, something that certainly does not assume the notion of truth or its relatives (and via coding, it can be fixed even in ML or, more generally, in any theory which has the resources of PRA). Tarski was in fact explicit on the need to move to 'a level one step higher' – to the metametatheory MML – while showing that a truth-definition satisfies Convention T (Tarski 1935, 195; Tarski 1936, 405); but he clearly considered translation as a rather unproblematic issue, and consequently said very little about it.

[9] I would like to thank three *Synthese* referees for valuable comments on an earlier version of this paper.

Department of Philosophy
P.O.Box 9
FIN-00014 University of Helsinki
Finland

E-mail: panu.raatikainen@helsinki.fi