MANUEL GARCÍA-CARPINTERO

WHAT IS A TARSKIAN DEFINITION OF TRUTH?*

(Received in revised form 16 February 1994)

Since the publication of Hartry Field's influential paper "Tarski's Theory of Truth" twenty years ago,[1] there has been an ongoing discussion about the philosophical import of Tarski's definition. Most of the arguments have aimed to play down that import, starting with that of Field himself. He interpreted Tarski as trying to provide a physicalistic reduction of semantic concepts like truth, and concluded that Tarski had partially failed. Robert Stalnaker and Scott Soames claimed then that Field should have obtained a stronger conclusion, namely that Tarski's failure in his allegedly intended physicalistic reduction was total.[2] From another front, Hilary Putnam argued for an even more sweeping thesis: "[a]s a philosophical account of truth, Tarski's theory fails as badly as it is possible for an account to fail."[3] Scott Soames followed suit also on this count, endorsing Putnam's argument in section III of the aforementioned paper; and John Etchemendy used a very similar argument to contend that "a Tarskian definition of truth [...] cannot possibly illuminate the semantic properties of the object language."[4]

These contentions, I shall argue, are based on several misunderstandings about the nature of a definition like Tarski's. Regarding the Putnam-Soames-Etchemendy argument, we must apply a distinction Frege found natural to draw – and is natural to draw anyway – between what he called "constructive" definitions properly so called, which are mere stipulations, and what he described as "analytic" definitions, definitions which aim to make explicit the meaning of a term already in use.[5] Although an analytic definition could conceivably look like a stipulative one, when, in providing it, we are claiming that the meaning the old term has is just what it would have had, had the users gathered around a table and agreed to use it in accordance with the explicit stipulation we are offering, the two kinds must be kept apart nonetheless. Frege himself claimed obscurely that the *content* of an analytic definition is very different from that of a stipu-

Philosophical Studies **82:** 113–144, 1996.

lative one, when he said that analytic definitions should be regarded as *axioms* instead of as proper definitions. But no matter how we characterize the difference, on the basis of a distinction like his we have enough material to dispose of the Putnam-Soames-Etchemendy argument. On the other hand, regarding the Field-Soames-Stalnaker criticism, I will contend that the true aims of a Tarskian truth definition are perfectly well fulfilled without Field's amendments.

In the first part of the paper (sections I–III) I shall be developing an account of what a Tarskian definition of truth is. As I call it 'Tarskian', I shall try to prove by means of quotations and other references to Tarski's work that the account does justice to his primary intentions. But my aim through the paper is mainly conceptual. Many people, beginning with Carnap and Popper, have seen in Tarski's definition an appealing explanation of the – or *a* – concept of **truth**. The principal guideline I have followed in designing the account I will provide is for it to live up to these expectations. There is nothing wrong thus in including elements opposed to some of the explicit declarations made by Tarski. I may still defend my calling the account 'a *Tarskian* definition of truth' – not merely as a tag the use of which admits only of a causal explanation, but with full descriptive force – if I can show that besides playing havoc with his declared purposes, those of Tarski's contentions that I will have to reject either contradict or do not follow from other, more highly ranked assertions also made by him.

After having explained what I take to be a Tarskian definition of truth, I will discuss in the second part of the paper (sections IV–VI) some of the issues raised in the above mentioned papers. I will develop my rebuttal of the Putnam-Soames-Etchemendy argument in sections IV and V, and I will discuss Field's criticism in section VI.

I

A *Tarskian definition of truth* is a definition of a predicate, let us say 'true_T', applying to the sentences (-type) of a language, L, which satisfies the following two conditions:

(i) The sense of 'true_T' agrees reasonably well with "one" of the concepts expressed by the ordinary predicate of truth

in natural language, namely, the "classical" or "Aristotelian." As a consequence, the extension of 'true_T' agrees with the extension of 'true', when this predicate is taken in the "classical" sense, and its domain restricted to the sentences of L.

(ii) The sense of 'true_T' is such that the semantic antinomies are not derivable.

There is not much to say in historical justification for the second condition. It is clear from Tarski's writing that one of his main purposes in constructing the definition was to avoid the semantic paradoxes, and nobody, to my knowledge, has disputed that. But before we go on to explain what is meant by 'classical', some philosophical comments are in order regarding the first condition. After all, it will play an essential part in my argument.

I might marshall a fair number of quotations from Tarski's writings in support of (i). In "The Semantic Conception of Truth and the Foundations of Semantics" (SCT in the following),[6] for instance, we read: "The desired definition does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary, it aims to catch hold of the actual meaning of an old notion" (SCT, 341), and afterwards "I do not have any doubts that our formulation does conform to the intuitive content of that of Aristotle" (SCT, 360). We can find similar assertions in the "Introduction" to "The Concept of Truth in Formalized Languages" (CTFL, in the following); see p. 153, and also §1, 155.[7] And also in "Truth and Proof" (TP, in the following), 63.[8]

Quotations, however, can be interpreted in different ways, and more to the point, their import can be ranked according to different orders. In the previously mentioned papers, Hartry Field and Scott Soames hold that Tarski's purpose in framing his definition was to help to further a reductionist program, namely, physicalism. They also have some quotations to offer. Thus, in "The Establishment of Scientific Semantics," Tarski writes, commenting on the possibility that an explicit definition of the semantic concepts may be not attainable, "it would arouse certain doubts from a general philosophical point of view. It seems to me that it would then be difficult to bring this method into harmony with the postulates of the unity of science

and of physicalism (since the concepts of semantics would be neither logical nor physical concepts)."[9] Moreover, they have an argument to show purportedly that Tarski's main aim could not have been what I have attributed to him in stating (i), because it is all too obvious that *that* aim could not be attained by his actual definition. I will discuss this argument in due course. Anyway, I will have to support in a more roundabout way than providing quotations my appraisal of condition (i) as being of the foremost importance in understanding what a Tarskian definition of truth is.

I shall discuss later the issue of physicalism too, and its connection with the problem of framing a truth-definition, Tarskian or other wise. For the moment, I want to emphasize that, to the extent that we can judge with any confidence on the basis of his explicit declarations, the issue was only ancillary concern to Tarski – if it was a concern of his at all. The text quoted in the preceding paragraph is a passing remark which can only be found in a paper written to be delivered for a physicalistically-minded audience (the International Congress of Scientific Philosophy, held in Paris in 1935), while the ones I have pointed out are passages in which Tarski either states his intentions or appraises their fulfillment once the definition has been presented, and that is so in every one of the relevant papers.

On the other hand, we have enough textual support to attribute to him a very different purpose. For several reasons (the philosophical muddles in which the discussion of the issue of truth usually ends up, the semantic paradoxes, the sheer appeal of the idea to an empiricist mind) many people in Tarski's intellectual circle had embraced the thesis that the concept of truth should be replaced in any "scientific" use by that of provability, or, better put, by *some* appropriate concept of provability. Tarski himself seems to have accepted the idea, as we can guess from some of his writings of the late twenties. But he soon rejected it, and we can find several declarations to that effect in writings from the early thirties on. The truth of a logical sentence, he now claims, must not be confused with its derivability in some system; and the same holds good regarding the truth of a mathematical sentence or the truth of a sentence in physics.[10]

Tarski's most frequently stated reason for rejecting the identification of truth and proof is that, according to the first concept, every sentence or its negation must be true – while it is not the

case that every sentence or its negation must be provable. If we had disputed this, pointing out for instance the problem of vagueness, or that of sentences with non-denoting singular terms, he could have answered, I believe, in this way: let us say, for the sake of this discussion, that a faulty sentence belonging to one of these kinds *does not express a proposition.* Now, if we consider only sentences which express propositions, it remains the case that both a sentence in this restricted class and its negation may be unprovable; but one of them (and only one) must be true. If we still disputed the issue, he would claim that this was what anyway *might* happen, according to an *intuitive* concept of truth; and that he was fully prepared to provide *precise* concepts of truth and provability on the basis of which he could *establish* that this was just what happened. (Using Gödel's Incompleteness Theorem.)

He has a less precise reason, thought. He emphasizes that the concept of truth guides us in devising proof procedures, and then in evaluating their strength and accuracy. The concept of truth has therefore a sort of *conceptual priority* over that of proof. He dwells on this point at the end of TP, and it is in this vein that he says in CTFL:

> The fact must also be taken in consideration that (in contrast to the concept of true sentence) the concept of provable sentence has a purely accidental character when applied to some deductive sciences, which is chiefly connected with the historical development of the science. (CTFL, 186.)

I think he would have extended the point to sciences other than the deductive ones, according to what he says in TP.

On the basis of all this, I maintain that we have more evidence to believe that one of Tarski's main aims in devising his definition was to legitimize (especially in the presence of the semantic paradoxes) "one" of the intuitive uses of 'truth', according to which it expresses a concept different from that of **proof**, than to believe that he was trying to achieve a physicalistic reduction of the semantic concept of **truth**, no matter what exactly such an achievement consists in. (Of course, it may well be that the two projects are not in conflict. But for purposes of interpretation, it matters which one is ranked first.) This is precisely what I claim in stating conditions (i) and (ii), as will be seen when I have indicated what the terms 'classical' and 'Aristotelian' in it mean.

In the previous paragraph, and also in the original formulation of condition (i), I have put 'one' into scare quotes, indicating that I leave open the possibility that the ordinary truth-predicate has more than one sense, but that I have some misgivings about it. I have tried to leave open that possibility only to keep faith with Tarski's explicit contentions in this respect. But I have used the scare quotes because I think he was being slightly disingenuous in leaving that possibility open. He was, I believe, merely trying to avoid disputes with the supporters of the so-called "correspondence," "coherence" and "pragmatic" theories of truth. I think that his considered opinion was that these "theories" are the product of a confusion, the confusion of *explaining a concept* with *giving general criteria for the application of a concept.*[11] (I include the "correspondence" theory here, obviously not meaning by it the "classical" or "Aristotelian" concept we are about to explain. I have in mind the "theory" according to which to be true is to correspond to an *epistemically privileged* subset of the facts, for instance the facts directly ascertainable by using the unaided senses. Of course, Tarski claimed that the truth-predicate he defined expresses a "correspondence" concept of truth, when we understand the term as meaning "Aristotelian.") He does not say it in so many words because, naturally enough, he does not want to deny the possibility of a well-defined concept of, say, "coherence" truth, and further because he does not want to engage in a futile dispute about which concept "is" the intuitive one. The only thing he needs to claim is that the concept his definition embodies does correspond reasonably well with the intuitive usage.

As I said before, Field and Soames have an argument to establish that (i) could not have been Tarski's aim – in spite of his explicit declarations to the contrary. Their point is that it is all too obvious that Tarski's actual definition fails to achieve that. I is so obvious that Tarski could not have failed to realize it, and how could he have taken on himself a task he knew too well he would not be able to fulfil? To Field, it is so obvious because the ordinary truth-predicate applies equally to the sentences of our language and to the sentences of any other language, with the same meaning; while Tarski's predicate is defined for a particular language. Soames' reason is slightly different, though related: the intuitive concept applies not only to sentences,

but also to propositions, and Tarski has nothing to say about this aspect of its use.[12]

To this, we must answer as follows. First, we do have a concept of truth which applies to sentences. (Or, to be precise, to sentences as uttered in a particular context.) And we need it; we want to evaluate as true or otherwise (utterances of) sentences. If *this* concept depends, as the Tarskian claims (and I will defend in due course that it does) on the semantic properties of the language, then, strictly speaking, this concept differs from language to language. Now, there are bound to be important similarities between the concepts of *true sentence of Spanish* and *true sentence of English* (as there will be between the concepts of *grammatical sentence of Spanish* and *grammatical sentence of English*). These similarities (which have to do with the extent to which both languages have similar means to express similar contents, tools with similar functions) can account for our idea of a general concept. They could be precise enough to allow us to define afterwards a general concept of *true sentence of some language*, as the related similarities could well provide for a concept of *grammatical sentence of some language*. At least, nothing seems to stand in the way of it. Second, armed with this concept of *true sentence of some language* we could easily explain the usage of 'true' as a predicate of propositions (provided we do not envisage propositions that cannot be expressed in any possible language, i.e., provided that propositions are some type of semantic value of possible sentences). And third, even if the usage of 'true' as a predicate of propositions were more frequent than its usage as a predicate of sentences (as uttered ...), a case can be made for preferring the order of explanation suggested here: we do not come across bare propositions. The notion of proposition must be explained to use precisely in the terms that we need, according to the Tarskian, to introduce the concept of truth in the first place.

All of this is obviously programmatic; I do not pretend for one moment to have made a convincing case that we could recover the ordinary use of 'true' as interlinguistic from a (genuine) Tarskian account of 'true' as a predicate of Spanish utterances. However, it is enough to discredit any argument, such as the previously considered ones, to the effect that *it is simply obvious* that Tarski's definition cannot be an attempt to explain the ordinary concept of truth. And

what is more to the point is that Tarski himself seems to have had in mind a program like the one just outlined. At least, this is the way I think we should understand section 2 of SCT, which ends with these words: "Of course, the fact that we are interested here primarily in the notion of truth for sentences does not exclude the possibility of a subsequent extension of this notion to other kinds of objects." (SCT, 342.) If it happens to be possible to capture the interlinguisticality of the ordinary truth predicate along the lines sketched above, a Tarskian account of 'true-in-L' might well stand a chance of satisfying condition (i). It is important that the very possibility of such a chance depends upon the fact that, contrary to the main assumption in the Putnam-Soames-Etchemendy argument, as I will try to show below, a Tarskian definition for L relies on semantic properties of L.

The sheer intelligibility of condition (i) forces us to acknowledge that the term 'formalized' in the title of CTFL must not be understood as if the concept there defined applied to uninterpreted languages. For the intuitive concept of truth, in whatever sense it is used, does not apply to uninterpreted sentences. Hence, the defined one must not have these entities in its domain, if we do not want to deviate sharply from the sense it is intended to grasp.

If some people have thought otherwise, it is because of a confusion caused by a very different use to which Tarskian definitions are ordinarily put, namely, that of giving the interpretation of an artificial language devised with some further purpose (usually that of illustrating how to define the concepts of logical truth and logical consequence). When used in these contexts, 'Tarskian definition of truth' does not express the concept we are endeavoring to explain, and simply means: *semantic interpretation for the language under consideration given by means of techniques like the ones introduced by Tarski to build his definition of truth.* Once again, Tarski himself makes no mistake:

> 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. (CTFL, 166–167)

In SCT, §20 he comes back to the point with a very apt metaphor: "the role of formalized languages in semantics can be roughly compared to that of isolated systems in physics." (SCT, 365.) And in TP he dwells on this way of putting the issue: "the only formalized languages that seem to be of real interest are those which are fragments of natural languages ... or those that can at least be adequately translated into natural languages." (TP, 68.) The point he is making in these texts gives us the correct sense in which Tarskian definitions can be said to be given for *formalized* languages. A Tarskian definition for any interesting language can only be constructed, for reasons we are about to consider, when the *logical syntax* of the language is known. To make explicit this logical syntax, we usually proceed by devising a "mock-up," a language with an explicit syntax built by abstracting out the complexities of the target language we think unrelated to our purpose. The sentences of this language are indeed as previously interpreted, as the sentences of the object language to which they correspond (perhaps "under a reading") were in the first place.

II

Now we come to explain the crucial concept in our first condition on Tarskian definitions of truth. We will say that a truth-predicate is *Aristotelian* or *classical* iff its definition includes all that is needed for it to satisfy *Convention T*, and *nothing else*.

Evidently I have a lot of clarifying to do. I shall start by "satisfying Convention T." I shall say that a truth predicate, Π, applied to utterances of sentences belonging to language L, *satisfies Convention T* iff it follows from the definition of Π, for every utterance of a sentence σ of L and context C, a sentence instantiating the schema (T):

(T) S (as uttered in C) is Π iff p,

where in the place of 'S' there is a standard name of σ, and in the place of p a sentence which *says the same as σ would have said when uttered in C, or gives the truth-conditions σ would have had when uttered in C*, or *expresses the proposition σ would have expressed when uttered in C*.

Several comments are in order. First, by *standard name* I mean a name built in the usual way by means of quotation-marks, or a "structural-descriptive" name constructed out of quotation-names of basic elements and an operation of concatenation. Second, it is commonplace nowadays that indexicals and other (ubiquitous) context-dependent words prevent us from characterizing Convention T in purely formal terms (i.e., replacing the stated condition with something like "where in the place of 'S' there is a standard name of σ, and in the place of p σ itself"). We want our truth definition, together with extra-linguistic information about the context of utterance, and the information that what John said was "I am hungry," to allow us to infer that what John said is true iff John is hungry. We want this, because, as I said, we do apply our truth concept to sentences as uttered in contexts. A purely formal condition would not give us this. (The former is not the only way, nor perhaps the best way, of formulating Convention T on the face of context-sensitivity. Probably it is better to settle for something like what Scott Weinstein proposes in "Truth and Demonstratives," p. 61.[13] But the precise formulation would be very complicated, and I do not want to clutter the discussion of the main issues with those details here. The previous formulation will do for the purposes at issue.)

Many people seem to feel that context-dependence is a picky nuisance which we do not need to care about here. After all, in the special case that L consists only of "eternal" sentences, and the language in which we give the definition contains L, the formal condition would do. This, I believe, is a mistake. We are interested here in the general problem of explaining what an account of our truth-concept looks like; we want to determine precisely the demands posed by such an account. We are not interested in the special problem of giving a truth-characterization for a particular language with unusual features. And this was Tarski's problem, too: he would have considered his methods a failure if they lacked application for the language of science. Granted, to supply a particular application of the account for the special case of a language consisting only of eternal sentences we can go along with something simpler. But it only begins to seem plausible to claim of an *Aristotelian* (in the defined sense) truth predicate that it is a *truth* predicate, intuitively speaking, when in the definition Convention T

is characterized in the non-formal, semantic way. If, in framing our definition, we were merely trying to satisfy the formal condition – only valid in a special case – it could easily happen that we neglected something essential to the general concept of truth we are trying to explain. (As a matter of fact, this is what would happen. We would miss the appeal we need to make to certain semantic properties of the language under consideration, to get an intuitively satisfactory account of truth for it.)

In the following text, Donald Davidson is making the mistake I have endeavored to reveal:

> Of course, a theory of truth is not treated as empirical if its adequacy is judged only in terms of the T-sentences it entails, and T-sentences are verified only by their form; this happens if we assume the object language is contained in the metalanguage.[14]

Even if the object-language is contained in the metalanguage, T-sentences are not verified only by their form. We can use a formal criterion, no matter which one, only when we have previously established that the formal criterion suffices to ensure the satisfaction of the relevant, semantic one. (And in many commonplace cases the formal criterion Davidson has in mind here will not do.) Therefore, the verification requires also to determine that certain semantic conditions obtain.

Context-sensitivity, to sum up, merely makes salient something that should be obvious once stated. The fact that a predicate correctly defined satisfies Convention T is a sufficient condition (or even a necessary one) for it to be a *truth*-predicate only to the extent that "to satisfy Convention T" implies satisfying the semantic condition I have stated under that name. Only then can it even begin to seem plausible that the satisfaction of Convention T is related to the intuitive question of truth. In certain favorable cases, a purely formal condition might be enough. However, from the point of view of capturing the truth concept, that is an accident. If, taking advantage of it, we abide by the formal characterization, we run the risk of coming to believe that less is involved in giving a truth definition than in fact is.

This is, by the way, the real reason why a definition of truth by means of substitutional quantification is not in general acceptable. (The definition could be: $\forall\sigma(\sigma$ is true_T in L iff $\Sigma\rho(\sigma = \text{`}\rho\text{'} \wedge \rho)$,

where 'Σ' represents the existential substitutional quantifier.) Tarski levels several objections against such a definition in CTFL, §1, all of which have satisfactory replies. The most important seems to be the charge that the semantic paradoxes could be derived. Tarski does not give an explicit argument to show why it is so. However, it has been pointed out that the argument Tarski most likely had in mind does not work, provided that the semantics for substitutional quantification is properly established.[15] I think that the proper objection to a definition by means of substitutional quantification is what can be inferred from the foregoing. In particular cases, such a definition is perfectly in order. But a definition by means of substitutional quantification, by itself, has not been shown to be a *truth*-definition.

Tarski was well aware that, in the general case, Convention T had to be characterised in semantic terms, and so he did. There is, however, a difference with my characterization. Where I say "... *expresses the proposition* ... " he says "... *has the same meaning* .... " The most reasonable way to draw the distinction between *meaning* and *truth-conditions* in the face of context-dependence justifies my election. We should play down this disparity, though. Our present concerns lie far away from Tarski's. The languages he had in mind lacked context-sensitivity; or, better put, the problems whose solution he wanted the truth definition for were unrelated to the question of context-sensitivity.

To end this section, I will justify the rider "and nothing else" in my characterization of an Aristotelian truth-predicate above. Obviously, no well-defined truth-concept for a language can disagree with Convention T; i.e., a well-defined truth-concept cannot entail, for some utterance of a sentence of L, a T-sentence whose right-hand side does not say the same as the sentence mentioned in the left-hand side says in the context of the utterance. A definition like that could still be materially correct, but the link with the intuitive concept of truth would have been severed. Now, what is characteristic of a Tarskian truth-concept is its *minimality*, a minimality that Tarski described in section 18 of SCT in terms of epistemic neutrality. Other philosophers have tried to capture that minimal character of Tarskian truth-concepts in different ways; thus, it has been said that the whole point of a Tarskian truth-predicate is to serve as a mere *disquotational* device, or to provide for *semantic ascent*.[16] A Tarskian truth-concept

agrees in this minimality with the so-called *redundancy* theory of truth. (The redundancy theory is even more minimal than Tarski's. The problem with it lies in that we lack a proper formulation of it, which is able to account for some of the most common uses of the *truth* concept, as Tarski himself pointed out in SCT, §17.) It is this *minimality* of a Tarskian truth-predicate what is aimed at with the rider "and nothing else" in the above characterization. A Tarskian definition will not impose, for instance, any additional epistemic conditions on the application of the truth-predicate.

Tarski saw this same minimality in Aristotle's classic definition, which explains both his previously quoted claim and my use of the term Aristotelian'. If we think that every statement can be cast in the form "... is" or "is not," as Aristotle (likely inspired by Greek grammar) seems to have thought, then we can see in Aristotle's famous dictum ("to say of what is, that it is, and of what is not, that it is not, is true; to say of what is not, that it is, and of what is, that it is not, is false") just another way of putting Convention T: "a saying about something, that it is (that it is not), is true, if it is (it it is not); is false, if it is not (if it is)." Or, briefly: "... is" is true iff ... is.[17]

Obviously, one of the main issues at stake is to determine just what is necessary to ensure the satisfaction of Convention T. We will come back to this important matter later, in discussing the Putnam-Soames-Etchemendy argument.

Tarski thought that the sense of any predicate defined by means of his methods, and that of the intuitive predicate, even when restricted to the sentences of some particular language, will never coincide. I have been describing his project as that of defining a predicate whose sense were that of the natural language truth-predicate (or one of them). But the truth is that, according to him, only a good approximation was possible. The reason is that he thought the intuitive predicate was "inconsistent."[18]

When we derive the Liar paradox,[19] we reason under the assumption that the intuitive truth concept includes the contention that every instance of schema T is true, even those that involve Liar sentences. After discovering that this idea leads to contradiction, we learn that we must abandon that assumption; Tarski's hierarchy is a way of doing it in a principled way. As I said, Tarski would have described this as a partial revision of the intuitive concept, but one which can-

not be applied to the ordinary truth-predicate. He thought that the "universal" character of natural language prevents us from viewing it as actually arranged in the hierarchical way his solution requires.

Although I do not think there is anything in principle absurd in the contention that an ordinary predicate is inconsistent, however, we do not need to say this, in this particular case. Tarski's reason is a good reason, I believe, to deny any solution to the Liar paradox consisting in seeing a Tarskian hierarchy in natural language.[20] (That is, consisting in claiming that, after all, the truth-predicate of English is not inconsistent, because there are many truth-predicates, arranged in a Tarskian-like hierarchical way.) We have independent reasons for the same conclusion, reasons nicely argued for by Saul Kripke in his "Outline of a Theory of Truth."[21] Nonetheless, Kripke himself offers in this same paper another solution to the Liar immune to Tarski's consideration about the "universality" of natural language, and, more importantly, compatible with the essentials of his explanation of truth.[22] Because of that, I will not follow Tarski in this regard; i.e., I will assume that the semantic paradoxes do not prevent us in any way from defining a truth-predicate capturing the sense of the ordinary truth-predicate, and not merely better or worse approximations.

III

With this interpretation as background, I will consider now the Putnam-Soames-Etchemendy argument (the *PSE argument* in the following). Putnam puts it nicely in this text:

> Now, pay close attention, please! This is just where, it seems to me, philosophers have been asleep at the opera for a long time! Since (2) [an instance of Convention T, with a Tarskian-style defined truth predicate, M. G.-C.] 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 assumption about the *use* of the 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. In fact, "true-in-L," as defined by Tarski, is a notion which involves only the primitive notions of L itself, as we said and as Tarski himself stressed. So, if L does not have notions which refer to the *use of linguistic expressions*, there is no way in which "true-in-L," or, rather, the notion to which Tarski gives that name, *could* involve the *use* of expressions in any way. 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.*

From the point of view of formal application of Tarski's theory in mathematical logic this doesn't matter, because all that a logician wants of a truth-definition is that it should capture the *extension* (denotation) of "true" as applied to L, not that it should capture the *sense* – the intuitive notion of truth (as restricted to L). But the concern of philosophy is precisely to discover what the intuitive notion of truth is. As a philosophical account of truth, Tarski's theory fails as badly as it is possible for an account to fail. 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.[23]

The argument Putnam is summarizing here can be stated as follows. Tarski's definition is just that, a *definition*, and besides a definition whose *definiens* does not include, apart from set-theoretic and syntactic terms, any other term than those already present in the object-language. Tarski himself stresses this; and, independently of his explicit claims, it is very important to his project for it to be so. For this is the way to achieve the end result of having a truth-predicate free from any suspicion that it could engender semantic contradictions. Suppose L is a language without semantic terms. If we define a truth-predicate for it using only the notions of L, plus set-theoretic and syntactic notions, then the only contradictions that could be derived from sentences containing this truth-predicate are bound to be set-theoretic, syntactic or in any event non-semantic contradictions. For if we have really *defined* the predicate, we ought to be able to eliminate it from any occurrence in favor of the terms used in the *definiens*. This defined predicate is then bound to be safe, as safe as the set-theoretic, syntactic or whatever other expressions already existing in L we used in the definition.

Now suppose we have a Tarskian truth-definition for a language (a fragment, or a mock-up, of English), one of whose sentences is the notorious 'snow is white'. Let us assume as above that the defined predicate is spelled 'true_T'. The definition, being a Tarskian one, will entail an appropriate T-sentence, say

(1) 'snow is white' is true_T in L if and only if snow is white.

This, mind you, will be a *logical consequence* of (i) the definition of 'true_T' plus (ii) certain facts. What facts are in (ii) will depend on the

nature of the definition. If L contained a finite number of sentences, a list-like definition like (2) (built in a meta-language that is a mixture of first-order logic and English under the assumption that L has just two sentences, 'snow is white' and 'grass is green') would do perfectly well, and the facts in question would be merely (tauto-)logical facts and equalities and inequalities among expression-types which can be discerned by mere inspection:

(2) $\forall\sigma(\sigma$ is true$_T$ in L iff σ = 'snow is white' and snow is white or σ = 'grass is green' and grass is green.)

If the definition belongs to a more interesting breed (unavoidably so, if L is an infinite language), then the facts in (ii) will include, besides logical and syntactic facts, set-theoretic facts and further definitional facts. (A Tarskian definition of any of the interesting varieties for a language with objectual quantification will be given by means of the usual detour through a previous definition of *satisfaction*, which in its turn will usually require previous definitions of the appropriate semantic relations for the names, the functional terms and the predicates of L. By the way, an interesting definition for a language which includes sentences like the ones we are considering here will be even more complicated, owing to the presence of mass terms like 'snow' and 'grass'.)

In any event, (1) will be at the most a logical consequence of definitions, logical truths, set-theoretic truths and truths contingent on trivial facts about spelling. This much we have *required* from a Tarskian definition of truth in the preceding sections. But logical consequences of this kind of facts are independent of semantic facts about the language under consideration. For these semantic facts are empirical facts, relative to the way the speakers use their language. The possible-worlds metaphor is a convenient way to make explicit this independence. Logical consequences of definitions, logical truths, set-theoretic truths and truths contingent on trivial facts about spelling hold in every possible world. Therefore, (1) holds in every possible world. Hence, in every possible world in which snow is white, the predicate we have defined, 'true$_T$', must apply to the sentence of L 'snow is white', *including those in which the conventions the users of* L *follow are slightly different and 'white' applies to black things.* To put it in a nutshell, whether or not 'true$_T$' applies

to 'snow is white' in L depends only on facts about the color of snow *and not on facts about the meaning of the expressions of* L. But this is at odds with our intuitions about the connection between the concept of *truth* and that of *meaning*. So extremely at odds that the defined predicate does not deserve to be considered a *truth* predicate at all.[24]

To many people, there seems to be something specious about this argument. But it is difficult to disclose what exactly this speciousness is. Martin Davies offers a recurring explanation: A Tarskian truth-definition is given for a particular language, on whose identity strict conditions have been imposed. A language in which 'white' means **black** is not the same language for which the definition was given in the first place, no matter how small the difference between them seems from the perspective of the ordinary, looser identity conditions for languages. It is not true, then, that 'true_T' would apply to 'snow is white' in a world like the one described. It would not, because the predicate's domain of application includes only sentences of L, and 'snow is white', in those possible circumstances, is not in their number.[25]

This rebuttal is essentially right, but, left on its own, it may seem a bit disappointing. This is because it does nothing to quell the doubts the argument may have raised about the ability of a Tarskian definition to account of the tight link between the concepts of truth and meaning, whose existence seems to be common ground. The problem is that the retort, as it stands, seems to originate in an all too easy stipulation, instead of drawing from some well-supported claim that a Tarskian definition can, and indeed does, catch hold of the intended connection between those concepts. I will try to flesh it out, having recourse to my previous account of the nature of a Tarskian definition.

IV

My strategy to show why the PSE argument is wrong will be to discuss first what I take to be an analogous case; i.e., a case such that an argument formally similar to PSE could be construed, while it is considerably easier to find out why the analogous argument would be specious. Then I will claim that PSE is fallacious for similar reasons.

In principle at least, a day may come when we are able to present the syntax of English in the form of an explicit definition of a predicate, say 'GRAMMATICAL$_t$', differing (from the point of view of the present discussion) only in complexity from one like the following:

(3) For all σ, σ is a GRAMMATICAL$_t$ sentence of L iff σ belongs to the smallest set Γ such that, for all ρ, $\rho \epsilon \Gamma$ iff (i) there is a NAME$_t$ of L, ν, and a PREDICATE$_t$ of L, π, such that ρ is the concatenation of π, '(', ν and ')', in that order – abbreviating: $\rho = \pi^{*}\text{`('}^{*}\nu^{*}\text{`)'}$ –, or (ii) there is a MONADIC LOGICAL CONSTANT$_t$ of L, χ, and a sentence in Γ such that $\rho = \chi^{*}\psi$, or (iii) there are sentences in $\Gamma\psi$ and υ and a DIADIC LOGICAL CONSTANT$_t$ of L ξ such that $\rho = \text{`('}^{*}\psi^{*}\xi^{*}\upsilon^{*}\text{`)'}$.

There are ancillary, list-like definitions of the non-syntactic terms used in the definition (3). They are gathered in (4):

(4) For all ν, ν is a NAME$_t$ of L iff ν = 'Mulhacén' or ν = 'Tübingen'; for all π, π is a PREDICATE$_t$ of L iff π = 'mountain' or π = 'city'; for all χ, χ is a MONADIC LOGICAL CONSTANT$_t$ of L iff χ = '¬'; for all ξ, ξ is a DIADIC LOGICAL CONSTANT$_t$ of L iff ξ = '∧' or ξ = '∨'.

Now, the following biconditional is a logical consequence of the definitions, plus set-theoretic and syntactic facts comparable to the ones used in the derivation of (1) from an interesting Tarskian definition:

(5) '¬mountain(Mulhacén)' is a GRAMMATICAL$_t$ sentence of L iff '¬mountain(Mulhacén)' = '¬'*'mountain'*'('*'Mulhacén'*')'.

As I said before, it seems to me perfectly possible that we are able to present the syntax of English in a similar fashion.[26] Of course, the definition has to be much more complicated. There will be intermediate categories, of different ranks. The rules will not be of the sort exhibited here, or not only of that sort at any rate. Perhaps the defined set will be a "fuzzy" one. Perhaps the definitions will appeal to more information than mere information about spelling. But none of this matters, as far as I can see. So let us assume that (3) and (4)

give an extensionally correct definition of the concept *grammatical sentence* of the language under study.

Now, an argument parallel to the PSE argument could be mounted against the claim that a definition of a predicate GRAMMATICAL$_t$ sentence of L like the one exemplified by (3) – together with the ancillary definitions in (4) – could constitute a good account of the concept *grammatical sentence of* L. This time the argument could run like this: Whether or not the concept *grammatical sentence of* L applies to a sentence is intuitively contingent upon facts about it that could be otherwise, facts about the way the speakers of the language use it. For instance, there are possible worlds in which 'mountain' is used as a diadic logical constant; in those possible worlds, the sentence '¬mountain(Mulhacén)' is not grammatical. But as (5) clearly shows, it follows from definitional, set-theoretic and other similarly non-contingent facts that the necessary and sufficient condition a sentence must satisfy for the defined predicate to apply to it obtains. Therefore, that a sentence is GRAMMATICAL$_t$ says nothing about the satisfaction of those contingent circumstances. As a matter of fact, things are even worse here than in the previous case of truth, for the necessary and sufficient condition a sentence must satisfy for the defined predicate to apply to it is not even contingent: it holds in every possible world whatsoever.

I take this as a *reductio ad absurdum* of the PSE argument. For (3) and (4) obviously grasp a possible grammaticality-concept; as I said, the analogous concept applying to the sentences of English should be characterizable by means of equally explicit definitions. But *what* is exactly the mistake in the argument?

The mistake consists in taking a definition like the one before as a stipulation, when it obviously aims to be something other than that. Any grammaticality concept for a given language which is at all like ours has to be complex: the grammaticality of a sentence must be necessarily dependent on some other facts. The reason is familiar. It is not that our grammaticality concept is *productive*, that it applies to an infinite number of sentences. That may be so, but the important reason is that it applies in a *systematic* way.[27] We could mention many symptoms of this. For one thing, when a new word is introduced in the language, it is thereby determined, without special stipulations, which combinations of it with old words are

sentences and which are not. We can assume, by way of example, that an explicit definition of the grammaticality concept like the one provided for GRAMMATICALITY$_t$ above attempts to make this apparent. In the final analysis, the definition attributes the grammaticality of '¬mountain(Mulhacén)' to its being the concatenation of certain shapes in certain order. (This is what (5) says.) But as I said, this is only "in the final analysis." To obtain that result, the definition works by classifying the shapes in several categories, and by establishing which combinations of categories constitute grammatical sentences. We would be leaving something essential out of the picture if we concentrated only on the end results of any application of the definition – like (5).

A definition like the one we are considering is not a mere stipulation, but what Frege called an "analytic" definition. It is more like a scientific theory, aiming at making explicit certain aspects of the meaning of a predicate already in use only implicitly known by those using it. The "observational" notions here are constituted by the observed systematicity in the way its users deploy the predicate *grammatical sentence* of the language under study. The purpose of the theory is to explain the systematicity observed in that property. The theoretical apparatus used by the theory to achieve its end is constituted by the categories it introduces and by the rules it establishes for putting together words belonging to several categories. The theory is explicitly presented as a definition, explicitly exhibiting the dependence of the application of the defined predicate on the theoretical ones. The empirical content of the theory comes from the assertion that the GRAMMATICAL$_t$ sentences of L are precisely the grammatical sentences of the language under study, already involved in the idea that the definition is an "analytic" one. To assert that involves contending that the theoretical apparatus employed by the theory actually *explains* why the sentences of L are grammatical. It involves contending that the old predicate applies to the sentences in virtue of facts like the ones which the application of the defined predicate depends on.

The link between the grammaticality of the sentences of a language and their syntactical structure is a necessary, constitutive one; a language composed of expressions belonging to the same morphological (or phonological) types but with a different syntax would

be a *different* language. Our "analytic" definition attempts to capture that link, that constitutive dependency relation in the case of L. Therefore, in counterfactual circumstances such that the same set of sentences are grammatical in virtue of having a different syntactical structure, the defined predicate does not apply to them. (Our definition-like theory is "empirical" in the sense that the same evidence – the same sentence-types being grammatical – is compatible with their having a different syntactical structure. But if the theory is correct, those sentences belong to other languages.) Hence, what is wrong in this syntactical variation of the PSE argument can be summed up the way Martin Davies does regarding the real PSE argument: the theory applies to a specific L, and what it would say about a language made up of the same words with different syntax is not determined. But there is no way here we could give the impression of providing an *ad hoc* answer: it is *precisely because* the theory displays the links between the properties of *grammatical sentence* and other properties that the possible language considered by the critic is beside the point.

These points apply as well to the actual PSE argument. The "empirical" claim of the theory is here already built in the characterization of a Tarskian definition of truth: as we have been at pains to show in the earlier sections of the paper, a Tarskian truth-definition it is the definition of a predicate *which agrees in sense with the intuitive one* to the extent that the definition allows the derivation of the right T-sentences. Thus, the claim that a definition is a Tarskian one can no longer be seen as a mere stipulation. The "observable" facts here are constituted by the T-sentences. The minimal claim made by the Tarskian is that all this sentences are analytically true, true in virtue of the meaning of the truth-concept. But there could be more interesting claims involved in the contention that the truth-definition captures the intuitive truth-concept for the language. The "theoretical" properties are those that are essential for the theory to satisfy its minimal empirical constraint, namely, to generate the right T-sentences. What theoretical properties they are depends on the type of the definition. In an uninteresting, list-like definition, they are correspondingly uninteresting: for instance, that the defined predicate applies to 'snow is white' if and only if snow is white. This was assumed as already known before (it is one of the "observ-

able" facts that, if the defined predicate is a truth predicate, it should apply to 'snow is white' if and only if snow is white), therefore it is uninteresting. However, if the theory is one of the interesting varieties (like a definition for a language in which we can project a first-order structure), the theoretical facts are correspondingly more interesting. Among them are included the logical form we attribute to the sentences (i.e. that they are composed in such-and-such a way of such-and-such expressions belonging to such-and-such semantic categories) and the contribution to the truth-conditions of the sentences in which they occur we assign to the so-called "logical constants" (quantifiers, connectives) in the relevant clauses. These are theoretical facts, because they are necessary for the theory to generate the right observable facts, the right T-sentences. In putting forward one of these more interesting definitions, the Tarskian claims that there is an explanatory link between some semantic facts regarding the language under study and the facts regarding the application of the truth-concept.

Again, these links are constitutive. The definition being an analytic one, it contends that a necessary connection exists between a sentence being true and its having whatever properties are required for the definition to specify the right conditions for it to be so. A language some of whose expressions lack some of those properties is beside the point to adjudicate the dispute over whether or not the definition is correct. In a possible world in which the sentence lacks some of those properties, the defined predicate neither applies nor does not apply to it. Not because of a convenient *ad hoc* stipulation by means of which we make things easier for us, but precisely because the definition *does* catch hold of some of those links between truth and meaning the critic contends it does not.[28]

Consider Soames' version of the argument I am criticizing:

> It is widely held that the meaning of a sentence is closely related to its truth conditions [...] Thus, many philosophers would accept arbitrary instances of (17) ...:
>
> (17) If 'S' had meant in *L* that *p*, then 'S' would have been true in *L* iff *p*.
>
> A natural demand growing out of this view is that substituting an adequate explication for 'true in *L*' in (17) should result in true sentences with contingent antecedents. [...] it is obvious that Tarski's definition does not satisfy this demand. For example, let 'Ws' be a sentence of *L* meaning that snow is white. using Tarski's definition of truth, we can produce the following counterpart of (17):

(17_T) If 'Ws' had meant in L that snow is black, then it would have been the case that snow was white iff snow was black.

[...] Th[is is] clearly not what the defender of (17)... has in mind. The reason [it isn't] is that Tarski's set-theoretic truth predicate doesn't impose any conditions ont he meanings of the sentence to which it applies. To be sure, Tarski wouldn't count any predicate T as a truth predicate unless $\ulcorner \alpha$ is $T \urcorner$ were materially equivalent to any metalanguage paraphrase of the object-language sentence named by α. On the basis of this, one might interpret Tarski as implicitly supposing that instances of (19) are necessary or a priori.

(19) If 'T' is a truth predicate for L, and 'S' means in L that p, then 'S' is T iff p.

However, this is quite different from maintaining that if 'T' in (20) is replaced with a truth predicate for L, then the resulting instances of the schema will be necessary or a priori:

(20) If 'S' means in L that p, then 'S' is T iff p.

It is this that the advocate of (17) ...demands and that Tarski appears not to provide.[29]

(17_T) is a simplification of the sentence "that would result from substituting Tarski's explicatum [...] for 'true in L' in (17) [...] The simplification [is] based on the fact that, where T is Tarski's explicatum, $\ulcorner$'Snow is white' is T$\urcorner$ and 'Snow is white' are necessarily equivalent (in the presence of elementary set-theory)."[30] We can now expose the fallacy in this argument. The simplification is not acceptable because, Tarski's definition being an "analytic," not a stipulative one, it establishes the necessary dependence of a sentence falling or otherwise under the truth-predicate on the facts about L relevant for the definition to deliver the right T-sentences. Therefore, the defined predicate neither applies nor does not apply to a sentence of the language in another possible world in which those facts are different. This is why the Tarskian has some claims to vindicate the necessity, aprioricity or even analiticity of (20), and not on the faulty grounds that (19) is necessary, a priori of analytic. His true ground rests on the dependence, which his theory claims to exists and aims to make explicit, of a truth-predicate applying to a sentence on the meaning of its semantically relevant parts. The fundamental Tarskian contention is that the facts on which, necessarily, whether or not the truth-predicate for a language applies to a sentence of the language

can be stated just as a by-product of stating the meaning-contribution of its semantically relevant parts.

There is a quick answer to a related criticism hinted at by Putnam in the text quoted at the beginning of the preceding section. He intimates that a Tarskian definition cannot possibly exhibit the intuitive connection between truth and meaning (i.e., use) because, to avoid the semantic paradoxes, it uses only the conceptual recourses already available in the object language; and those frequently do not express semantic properties, particularly properties describing the use the speakers make of their language. My argument has been that a definition belonging to any of the interesting brands does use theoretical properties related to meaning, for instance, *being an expression which applies* (i.e., not one which *refers*). These properties do not explicitly concern the use the speakers make of the language. But this is as it should be. Biological facts depend presumably on chemical facts, but it is legitimate for biology to use autonomous theoretical properties, in so far as they might be shown to be reducible to, or, more sensibly, supervenient on, chemical properties. It is the same with semantics. Part of the semantic task is to show how some semantic properties depend on others. (In the case that concerns us, how the concept of truth depends on some semantic properties of the language to which it applies.) These explanations are O.K. from a physicalistic viewpoint to the extend that the ultimately explanatory semantic properties depend in the required way, whatever it is, on more basic properties. Putnam would need an argument establishing that there is no hope of showing that specific semantic properties like the one previously mentioned cannot be proven to be supervenient on sociological and/or psychological properties. Perhaps that might be done, but nothing in the argument we are considering here gives us any hint about how.

My characterization of a Tarskian definition of truth makes use of the intuitive concept of truth; and that use is obviously essential to my argument, as the reader will have realized. This can be found objectionable: Was not Tarski's project to provide an appropriate substitute for the intuitive concept of truth, which he and his contemporaries considered suspect? This has been contended, in almost so many words, by John Etchemendy:

So far I have emphasized that Tarski's aim was not to give a semantic theory of the object language, but to define a predicate which, in virtue of its satisfying Convention T, could safely be used to express what would otherwise require the use of a concept whose consistency seemed questionable. What makes it hard to keep this difference firmly in view is the fact that, thanks to the techniques Tarski uses in his definition, the claim that all and only the true sentences of the language are members of the defined set takes on genuine semantic import. Yet such claims are most emphatically not part of Tarski's project, but in an obvious sense conflict with it, involving as they do the uneliminated use of a notion of truth.[31]

The "claim that all and only the true sentences of the language are members of the defined set" is what I have considered the "empirical claim" included in the contention that a given definition is a Tarskian definition; and just for the reasons Etchemendy intimates here I have concluded that Tarskian definitions have usually "genuine semantic import." Now, Etchemendy reasons against my characterization are that "such claims in an obvious sense conflict with Tarski's project." Why is that so? Because they involve "an uneliminated use of a notion of truth." The idea seems to be that to be as safe as possible from the threat of any antinomy, we should not use any other truth predicate than an eliminable one.

This may be sound policy, but the argument it allegedly supports is not valid. The defined predicate has been proven free from contradiction, for it is eliminable. But just because that, when we claim that the defined predicate is a truth predicate, that it catches hold of the sense (*one* sense) of the predicate 'true', *to the extent that our claim is correct*, we are using in making it a predicate as safe as the defined one has been proven to be.[32]

V

Field's views are so familiar by now that a cursory outline will be enough before discussing them. As I said before, he claims that Tarski's true goal was to achieve a physicalistic reduction (or at least a physicalistic explanation) of a semantic notion, *truth*. He is sympathetic to that aim, but thinks that Tarski failed to attain it. A typical Tarskian-like definition for a language with a first-order logical structure allows us to eliminate the defined truth-predicate in favor of notions already present in the language. To the extent that these notions are physicalistically acceptable (and so are the

set-theoretic and syntactic notions we use in addition), it can be seen as offering an extensionally correct reduction. But it does not give us what we expect from a truly physicalistic reduction, because it is not *explanatory*. The problem lies in the list-like clauses for the primitive non-logical terms, namely the ancillary definitions of *denotation* for the primitive singular terms and the primitive functional terms, and of *application* for the primitive predicates. These clauses do not tell us, in non-semantic terms, *why* the primitive non-logical terms contribute as they do to the being true or otherwise of the sentences in which they occur.

Field outlines an alternative. According to it, we would have as definition of the technical term DENOTATION, for any primitive singular term ν, something like: ν DENOTES what it denotes. As definition of the technical term APPLICATION for any primitive predicate π we would have something like: π APPLIES to something if and only if π applies to it. Obviously, a definition like this would generate T-sentences whose right-hand sides would not be free of semantic terms: 'mountain(Mulhacén)' is TRUE if and only if 'mountain' applies to what 'Mulhacén' denotes. To attain a full but really explanatory physicalistic reduction, Field trusts general explanations of the relations *denoting* and *applying* in the terms of what are generically called *causal* (or *historical*) *chain theories of reference*. What we could get in the end that way would be something like: 'mountain(Mulhacén)' is TRUE when uttered in C if and only if what is at the end of the causal chain leading to the use of 'Mulhacén' in C has the property at the end of the causal chain leading to the use of 'mountain' in C. This is general; but, as I said before, probably we need to characterize convention T as merely asking for this kind of general T-sentences anyway, owing to the presence of indexicals. And it generalizes where the original Tarskian definition does not. If the language were augmented with a new primitive singular term or predicate, we would need to modify our Tarskian-like definition of TRUE; but we would not need to modify a definition along the lines outlined by Field in such a case.

Field seems satisfied with the explanatory adequacy (for physicalistic purposes) or the clauses for "logical expressions" in typical Tarskian definitions. Now, as Robert Stalnaker and Scott Soames

have pointed out, there are reasons to doubt this. Soames put it this way:

> Field's objection to this [a Tarskian clause for a predicate, 'R', M. G-C.] is that although Tarski's definitions correctly *report* that 'R' applies to different things in the two languages [Soames is considering two Tarskian truth-definitions for two different languages in which 'R' applies to different sets of things], they don't *explain* how this differences arises from the way in which speakers of the two languages use the predicate. What Field fails to point out is that exactly the same objection can be brought against Tarski's treatment of logical vocabulary and syntax in the recursive part of his definition. (...) [Now Soames gives two truth-definition clauses for two apparently similar languages that differ in that the clause for '∨' in one is the usual clause for disjunction, whereas the clause for the same sign in the other language is the usual one for conjunction.] Owing to this difference, sentences containing '∨' will have different truth-conditions in the two languages. In order to satisfy Field's requirements on reduction, it is not enough for a truth characterization to report such differences. Rather, such differences must be explained in terms of the manner in which speakers of the two languages treat '∨'. Since Tarski's truth definitions don't say anything about this, their recursive clauses should be just as objectionable to the physicalist as the base clauses.[33]

This criticism seems to me well-taken. But neither it, nor Field's own, affects Tarski's definition, as I have proposed to understand it. For Tarski's true aims (validating the Aristotelian understanding of the ordinary truth-concept, by proving it eliminable and therefore free from contradiction), as opposed to the ones attributed to him by Field and Soames (furthering physicalism) a list-like definition (when it is possible) or an ordinary one, involving list-like definitions of the "primitive denotation" relations, is perfectly in order.

Nevertheless, I would like to add some remarks on physicalism and truth theories. I, too, feel strongly sympathetic towards physicalism. I think that we need physicalistic reductions for those "macroscopic" properties that we regard as causally efficacious. We need them to *validate* the causal efficacy of those properties. Reduction is achieved, though, by the discovery of contingent but nomic laws, relating the macro-properties to some class of lower level properties. (As a matter of fact – owing, among other things, to the multiple realizability of macro-properties by lower level properties – it is more advisable to settle for (nomic) *supervenience* than aiming for reduction.) Be it as it may, it is not up to us, philosophers, to establish the physicalistic pedigree of a property considered as suspect. Of course, what we can do is something general in character: to show

that conceptual arguments to the effect that a class of properties are not amenable to physicalistic reduction are without basis, or, more positively, to indicate in a general way how a class of properties could be physicalistically reduced. I take it that this is what Field intends to do, although I am not quite sure.

The property of *truth* is one of those in need of physicalistic reduction, also according to my own viewpoint. This is so because we want to use it in causal explanations.[34] But to my understanding, a moderate minimalist account of *truth*, like the one a Tarskian definition gives us for specific languages, has all we need to satisfy our reasonable physicalistic yearnings *regarding truth itself*. Of course, we still need a physicalistic account of the properties we rely on to get the right T-sentences; a physicalistic account both of why $\ulcorner \alpha \vee \beta \urcorner$ is true if and only if α is true of β is, and also of why 'mountain' applies to something if and only if it is a mountain. But I think we have made a strategically very important move forward in having displaced the physicalistic worry regarding truth towards physicalistic worries regarding these other properties. We do not know how to start accounting in more basic terms for the truth-conditions of sentences, while we have some ideas regarding how to account in more basic terms for those other properties. Part of the reason is that there are many sentences with different truth-conditions, but not so many basic constituents of them.

This also accounts for my reasons to disagree with the extreme minimalism (as opposed to the *moderate* minimalist the Tarskian holds) defended by Paul Horwich in his book *Truth*.[35] A Tarskian definition for any interesting language requires, according to my interpretation, a good amount of semantic theorizing for that language. Particularly, it requires a good part of what traditionally has been thought as a theory of reference for that language, i.e., of an account of how the content or truth-conditions of sentences depends systematically on some semantic values of their parts. Horwich's minimalism would do without it. But I do not think it does. The main problem is that the theory Horwich propounds is extremely ill-specified. He encloses a sentence between angles to refer to the proposition it expresses, and characterizes the theory in this way: "The axioms of the theory are ... all the propositions whose structure is: $\langle\langle p\rangle$ is true iff $p\rangle$" (Op. cit., 18). This does not tell us how to

get the corresponding axiom for a proposition given otherwise than by means of a sentence that expresses it; nor how to get an axiom when the proposition is given by a sentence not fully expressing it, because the sentence contains indexicals or other context-dependent devices. It does not tell us either whether, say, $\langle\langle 2 + 2 = 4\rangle$ is true iff $3^2 = 9\rangle$ is an axiom of the theory. The author refuses to give us his identity conditions for propositions; but if propositions are sets of possible worlds, the former should be an axiom, and if we do not want axioms like that, we incur in an obligation to individuate propositions otherwise, and to say how.[36] This unability to rightly present the theory is not a minor defect, but on the contrary is related to the main difference between moderate and radical minimalism: It is precisely because the Tarskian wants to give a precisely specified theory that he needs to have resort to some of the semantic properties of the object-language.

Horwich claims that he can give a theory as minimal as the one outlined for *utterances* instead of for propositions; but what he actually offers, in section 34 of the book, are two accounts. One is in fact not a theory, but a *condition* of adequacy for theories similar to a version of Convention T I myself discussed above – using 'translation' where I used 'proposition'. The other is truly minimal, but it does not satisfy intuitive constraints that Horwich himself places on any acceptable account of truth: in this version, by definition, 'truth' applies only to the utterances in the right-hand side of instances of Schema T; there is thus no chance that we could apply the theory to infer consequences of premises establishing that some utterance of someone is true, for the defined 'true' simply does not apply to those utterances. I very much doubt that a theory as minimal as Horwich intends, not having recourse to a specification of the reference of the subsentential parts, could be properly formulated.[37]

NOTES

* Among the many people who have helped me in articulating the ideas in this paper, I would like to express my gratitude to John Etchemendy, Hartry Field, Mario Gómez, James Higginbotham, Paul Horwich, Ramón Jansana, Ignacio Jané, Josep Maciá, Begoña Navarrete, Stephen Schiffer and Robert Stalnaker. The paper has greatly improved both in truthfulness and in clarity because of their comments, criticisms and suggestions; any remaining falsehood or confusion, however, should most probably be blamed on my stubbornness. The first version

of the paper was completed while I was a visiting scholar to the Center for the study of Language and Information, Stanford University. I thank the CSLI for the warm and encouraging atmosphere I found there, and the Spanish DGICYT for the grant that made possible my stay at the CSLI. This paper is part of the research project PS94-0244 funded by the DGICYT.

[1] *Journal of Philosophy*, LXIX (1972): 347–375.

[2] See Robert Stalnaker, *Inquiry*, Cambridge, Mass.: Cambridge University Press, 1984, chapter 2, and Scott Soames, "What is a Theory of Truth?," *Journal of Philosophy*, LXXXI (1984): 411–429.

[3] "A Comparison of Something with Something Else," *New Literary History* XVII (1985): 61–79, p. 64.

[4] John Etchemendy, "Tarski on Truth and Logical Consequence," *Journal of Symbolic Logic*, 53 (1988): 51–79, p. 56.

[5] See Gottlob Frege, "Logik in der Mathematik," *Nachgelassene Schriften*, H. Hermes, F. Kambartel and F. Kaulbach, eds. (Hamburg: Felix Meiner Verlag, 1969) pp. 219–270.

[6] Alfred Tarski, "The Semantic Conception of truth and the Foundations of Semantics," *Philosophy and Phenomenological Research*, IV, 1944, pp. 341–376.

[7] Alfred Tarski, "The Concept of Truth in Formalized Languages," in A. Tarski, *Logic, Semantics, Metamathematics*, J. Corcoran, ed. (Indianapolis: Hackett Pub. Co., 1983).

[8] Alfred Tarski, "Truth and Proof," *Scientific American*, 220 (1967), 63–77.

[9] Alfred Tarski, "The Establishment of Scientific Semantics," in A. Tarski, *Logic, Semantics, Metamathematics*, J. Corcoran, ed. (Indianapolis: Hackett Pub. Co., 1983).

[10] See CTFL, 186, text and note; 237–238, text and note 2, and 254, note. See also SCT, §12, 354, and §22, 368. The explicit intention of the popular paper TP is to present to a general public his results on truth. What he chooses to do (from the very title on) is more than that: he strives to make clear the relevance of his definition to the distinction between truth and proof.

[11] In several passages he emphasizes that his definition does not provide general criteria for the application of the predicate, but that is quite all right, because accounting for a concept does not need to involve giving criteria for its application. See CTFL, 197 and note, and SCT, §18, 361–362, and §19, 363–364.

[12] See Hartry Field, op. cit., §II, and S. Soames, op. cit., pp. 411–412. Actually, from the historian's point of view Field's course of argument is rather extraordinary. He interprets Tarski as if his aim were not what he explicitly and repeatedly declares it is, but something else (to attempt a physicalistic reduction) he only collaterally claims his definition can provide, ... only to show then (by means of a quite convincing argument, I must say) why Tarski fails miserably to attain that alleged aim. If I manage to show that Tarski actually succeeded in his explicitly declared aim, this can be taken as yet another indication that the problem of getting a physicalistic reduction is more a concern of Field than one of Tarski.

[13] In Davidson and Harman (eds.), *The Logic of Grammar*, Encino, California: Dickenson, 1975.

[14] Donald Davidson, "In Defense of Convention T," in D. Davidson, *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, 1980, p. 73. I think that Quine hides important problems of his philosophy of language under

his familiar contention that we take our sentences "at face value," by which he seems to understand that we apply to them a purely disquotational truth-predicate. A purely disquotational truth-predicate would then be one that satisfies Convention T characterized in a purely formal way. I do not believe we have such a predicate; and such a predicate would not anyway be a *truth*-predicate, for the reasons pointed out in the main text.

[15] See Saul Kripke, "Is There a Problem about Substitutional Quantification?", in G. Evans and J. McDowell (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Oxford University Press, 1976, pp. 325–419; see especially pp. 367–368.

[16] Notoriously, W. V. O. Quine. See his *Philosophy of Logic*, Cambridge, Mass.: Harvard University Press, 1970, pp. 10–13, and the entry 'Truth' in his *Quiddities*, Cambridge, Mass; Harvard University Press, 1987.

[17] On the assumptions that "... is" is false iff ... is not, and that the metalanguage includes the object language.

[18] When the sense of a predicate can be given by specifying a set of open sentences, it makes good sense to characterize the predicate as *inconsistent*. We mean that logically nothing can satisfy the set.

[19] I will be assuming that a *paradox* is an argument that from plausible premises leads to a contradiction.

[20] By a *solution* to a paradox I mean here an account according to which the contradiction is merely apparent: some premise in the argument is as a matter of fact false, although for whatever well-entrenched motives we tend to reckon it to be true.

[21] *Journal of Philosophy*, LXXII (1975), pp. 690–716.

[22] In particular, compatible with everything I will have to say below on Tarskian definitions of truth.

[23] Hilary Putnam, op. cit., pp. 63–64. Etchemendy and Soames make the same point in sections 1.2 and III, respectively, of the above-mentioned papers.

[24] The same point can be made considering the intuitive connection between *truth* and *understanding*, and the lack thereof in the case of the defined predicate. See S. Soames, op. cit., §III for illustration.

[25] Martin Davies, *Meaning, Quantification, Necessity*, London: Routledge and Kegan, 1981, 28.

[26] I am speaking here about the *syntax* of English, not about the *knowledge* the speakers have of it. (Unless by 'knowledge' you mean the *object* of the state of knowing, and not the state itself, in which case we do not really disagree.) I think this distinction can, and should be drawn, in spite of Chomskyan arguments to the contrary. In the last analysis, the syntax of a language is a system of *complex conventions*; a syntactic theory displays that complexity. In a similar vein, when I speak here of *concepts* I mean *the meaning of a predicate*, not the knowledge its users have of it. These two pairs are strongly related, their relationship is not identity.

[27] For the distinction between the two issues, see for instance the appendix to Jerry Fodor, *Psychosemantics*, Cambridge, Mass.: MIT Press, 1987.

[28] This does point to an interesting use of the construction of Tarskian theories as a heuristic tool in semantics.

[29] Soames, op. cit., pp. 422–424.

[30] Ibid., 423.

[31] John Etchemendy, op. cit., p. 60.

[32] Tarski's contention that the English truth-predicate is "contradictory" could raise some doubts in this regard. This is another reason not to follow him on this matter as I suggested that we should not in section II above.
My disagreement with Etchemendy on the significance of Tarskian truth definitions accounts also for my disagreement with him on his dim evaluation of Tarskian, model-theoretic, analyses of the logical properties – *logical truth* and *logical consequence* – in his *The Concept of Logical Consequence*, Cambridge, Mass.: Harvard University Press, 1990. See my "A Defense of the Model-Theoretic Account of the Logical Properties," *Notre-Dame Journal of Formal Logic*, 34 (1993): 107–131.
[33] Scott Soames, op. cit., 419f.
[34] See Hartry Field, "The Deflationary Conception of Truth," in G. MacDonald and C. Wright (eds.), *Fact, Science and Morality*, Oxford: Basil Blackwell, 1986.
[35] Oxford: Basil Blackwell, 1990.
[36] Actually, Paul Horwich told me in personal conversation that he intended the axioms as specified by putting the same sentence in both 'p' positions. This is in fact to say that the axioms are not propositions, but propositions "under a linguistic guise."
[37] I owe some of the points in the last two paragraphs to my colleague, Ignacio Jané.

Departamento de Lógica, Historia y Filosofía de la Ciencia,
Universidad de Barcelona
08028 Barcelona
Spain