Disputatio 16, May 2004

Tarskian truth and the two provinces of semantics[1]

Ricardo Santos
New University of Lisbon

Abstract
In this paper, I argue that the cleavage between the theory of reference and the theory of meaning, which under the influence of Quine has dominated a large part of the philosophy of language of the last fifty years, is based on a misrepresentation of Tarski's achievement and on an overestimation of the scope and value of disquotation. In particular, I show that, if we accept Davidson's critique of disquotation, the same kind of reasons that Quine offered in opposition to the Carnapian theory of meaning also apply, mutatis mutandis, to the Tarskian theory of reference.

When Tarski presented his theory of truth, he wanted to achieve several goals. Among them there were the more philosophical ones of solving the so-called classical problem of truth and of establishing the legitimacy of semantic concepts, thus laying the foundations of semantics as a scientific discipline in its own right. Carnap and Popper are usually mentioned as revealing witnesses of Tarski's success as far as these goals are concerned. But Quine, who Tarski even thought of writing a book

[1] Work for this essay was supported by a postdoctoral fellowship awarded by the Portuguese Foundation for Science and Technology, and was included in the activities of the research program *The Explanation of Human Interpretation*, developed at the Institute for the Philosophy of Language (New University of Lisbon). A previous version of the essay was presented, in June 2002, at the 4th European Congress for Analytic Philosophy, at Lund University, Sweden. My greatest debt is to João Sàágua for initial discussion of these issues. I am also grateful to Fernando Ferreira and Dina Mendonça for helpful comments on this version.

together with[2], is a no less significant witness, most of all by the way he tried to show that this success does not extend to the whole of semantics, but only to that part which he proposed to call the 'theory of reference', as opposed to the more suspicious 'theory of meaning'. For Quine there is no philosophical problem of truth to be solved after Tarski[3], as there is also no place for any doubt concerning the legitimacy of the concepts from the theory of reference, even after their relativity has been recognised. But regarding the concepts of meaning, synonymy, analyticity, necessity, etc., Quine holds that the sceptical attitude that used to affect all semantics before Tarski should be kept. These 'two provinces' of semantics are, in his words, 'so fundamentally distinct as not to deserve a joint appellation at all'[4]. In this separation, Quine's assessment of Tarski's work plays a role, which is no less important than the one played by his own criticism of the notions of meaning and analyticity for their lack of intelligibility.

Few philosophers, though, would share today Quine's optimistic view of Tarski's philosophical achievements. On the one hand, as for the legitimacy of semantics, Field's arguments are known for claiming that Tarski only reduced the concept of truth to other more primitive semantic concepts[5]; moreover, they were complemented by the arguments of those who showed that not even with that partial reduction should he be credited[6]. On the other hand, as for the philosophical problem of truth, many authors agree today that Tarski didn't capture or didn't provide a satisfactory analysis of our intuitive notion of truth. Putnam, for instance, believes that 'As a philosophical account of truth, Tarski's theory fails as badly as it is possible for an account to fail'[7]. And Davidson, for many

[2] See Quine (1985) 190.

[3] As an expression of this opinion, see the statements in Quine (1992) 82: 'We understand what it is for the sentence 'Snow is white' to be true as clearly as we understand what it is for snow to be white. Evidently one who puzzles over the adjective 'true' should puzzle rather over the sentences to which he ascribes it. 'True' is transparent.'

[4] Quine (1980) 130.

[5] See Field (2001) 3-26.

[6] See, *e.g.*, Soames (1999) 110-112.

[7] Putnam (1994) 333.

people's surprise, though he is not so radical, has also concluded that Tarski 'did not capture essential aspects of the concept of truth'[8].

But what are the reasons for this failure? Which aspect of Tarski's theory should be held responsible for it? When we ask these questions, opinions start to diverge. Let's concentrate on Putnam's argument, which has also been presented, in more detail, by Etchemendy.

What they both point out as problematic is the 'tautologous' character of the T-sentences that one can infer from a Tarskian truth definition. A sentence as

(1) 'Der Mond ist blau' is true in German if and only if the moon is blue

should be very informative, expressing as it seems a contingent truth, crucially dependent on the ways German speakers use the words of their language. To know that the moon being blue is what is required for that German sentence to be true, to know that those are the truth conditions of the sentence, is to have an important information about the German language – so important that, according to Davidson (and also Frege and Wittgenstein), if we had an analogous information for all the sentences of the language, we could be said to understand the language. It was that, in fact, what led Davidson to defend in 'Truth and Meaning' that Tarski's definitions of truth could work as theories of meaning for the languages they relate to[9].

But if sentence (1) is a T-sentence inferred from a Tarskian definition, in other words, if the truth predicate occurring in it is a Tarskian defined predicate, that informativeness is illusory. We only read that semantic information into the sentence because we load the Tarskian predicate with a content it doesn't have. In fact, being a predicate that has been explicitly defined – and that means: defined in such a way that its elimination from all contexts becomes possible –, its meaning is just that which the definition gives it. But if, following this principle, we substitute in (1), for the predicate 'is true in German', its *definiens*, what we will end up with is a trivial logical truth (or, more exactly, a truth of logic, syntax and set theory). That is to say, we get a sentence (1*) that, unlike what we

[8] Davidson (1990) 288.
[9] See Davidson (1984) 24.

thought about (1), is true regardless of the meaning of 'Der Mond ist blau'. Noticing this, Putnam concludes that the property expressed by Tarski's predicate may be very interesting for logico-mathematical purposes, but it certainly cannot be the property of being true (as we ordinarily conceive it). For truth could never be identified with a property that that sentence has in all possible worlds in which the moon is blue, including worlds in which it means very different things such as, for instance, that the moon is green. In Putnam's view, even relative to a language, the property that Tarski defined 'just isn't truth at all'[10].

When Etchemendy, on his turn, shows that Tarski's T-sentences are logical truths, he has a different primary goal: he wants to prove that those sentences cannot serve the semantic purpose desired by Davidson. However, Etchemendy has also a second argument showing that, if we substitute in T-sentences, for the defined predicate of Tarskian truth, a predicate expressing what he calls 'a primitive notion of truth', we end up having precisely what Davidson needs to compose a theory of meaning[11]. Also, it seems to easily follow from here that that primitive notion is not captured by Tarski's definition. Etchemendy doesn't share Putnam's pessimistic conclusion only because he refuses to identify that primitive notion with what Putnam has called 'the intuitive notion of truth'.

But should this remark, that T-sentences are logical truths, cause so much surprise? In his book *Representation and Reality*, Putnam reports the way he already in the fifties confronted Carnap with this remark, having received an unsatisfactory reply according to which the feeling of strangeness would vanish as soon as we look to languages as abstract objects identified by their semantic rules[12]. And, in the same book, Putnam also mentions how Quine, confronted with the same fact, found it '*very* counterintuitive'[13]. However, I find it reasonable to suppose that Tarski himself would not be very surprised. For, after all, it is precisely to make T-sentences come out as logical consequences of the truth definition that he frames this last one. As he wrote: 'Not much more in principle is to be demanded of a general definition of true sentence than that it should [...]

[10] Putnam (1994) 333.

[11] See Etchemendy (1988) 59.

[12] See Putnam (1988) 63.

[13] Putnam (1988) 132n7.

include all partial definitions [*i.e.* all the T-sentences] [...] as special cases; that it should be, so to speak, their logical product.'[14]

Tarski thought that, to be satisfactory, a definition of truth must be formally correct and materially adequate. Formal correction imposes the usual conditions of eliminability of the defined term and non-creativity of the definition. To these conditions Tarski adds his 'reductionist' wish that no undefined semantic term be used in the definition. Meeting these formal conditions will guarantee that any sentence in which the truth predicate occurs – as it happens in T-sentences – can be transcribed in a sentence in primitive vocabulary exempt of any semantic term. But nothing until now guarantees the provability of such transcriptions. Transcribing T-sentences in primitive notation might give rise to true but not provable sentences of the metatheory. This, seemingly, is what Putnam would find proper: that the axioms of logic and set theory (and, maybe, of syntax[15]) be not sufficient to prove T-sentences. Tarski deviated from this when he required T-sentences to be consequences of the definition. For, according to non-creativity, if they are provable from the definition, they are also provable directly from the axioms of the metatheory. And why did Tarski add this last requirement? It was to guarantee the material adequacy of the definition, to guarantee that the defined predicate does not 'denote a novel notion', but rather 'catch[es] hold of the actual meaning of an old notion'[16]. In Tarski's view, had the definition not allowed the proving of the T-sentences, we would have no means to know that the predicate it introduces is a truth predicate. There is, therefore, a great irony in this: it was to guarantee the adequacy of the definition that Tarski required T-sentences to be, not only true, but also provable; and it was exactly because of this requirement that he transformed T-sentences into logical truths, and by doing that he, according to Putnam, failed that very same goal of adequacy.

Davidson also wasn't very surprised with the tautologous character of T-sentences. In his view, that is only a predictable effect of the enumerative nature already pointed out in Tarski's definitions by several authors.

[14] Tarski (1983) 187.

[15] But, as it is well known, this can be arithmetized, *i.e.* reduced to (or simulated in) set theory.

[16] Tarski (1944) 341.

Davidson gives it the following explanation: 'if the extension of a predicate is defined by listing the things to which it applies, applying the predicate to an item on the list will yield a statement equivalent to a logical truth'[17]. This is a valuable remark, though its attribution to Tarski is not entirely correct. It is not correct because, even in the simplest case of a language with only a finite number of sentences, Tarski doesn't define truth by listing the true sentences of the language. For Tarski rightly supposes that we might not know which sentences are true. What we certainly know is, for each sentence, what has to be the case for it to be true – and this knowledge, which a T-sentence expresses, we already have it before the definition, and it's also with it that we then test whether the definition is adequate. We need, then, to distinguish the status T-sentences have *before* and *after* the definition, being pretty evident that they aren't logical truths before the definition. Now, it's exactly here that Davidson's remark may become of great help, by calling our attention to one way that a statement possibly with empirical content (*viz.* the application of a predicate to an object) may be transformed into a logical truth (*viz.* by defining the predicate as a predicate that, among other things, applies to that object). Definition by list, though, isn't the only way of doing that. We could also, for instance, turn the sentence 'Snow is white' into a logical truth, if we define 'snow' as 'such and such white substance'. Given this definition, that sentence would become true even in a counterfactual situation in which snow was blue. And what does this show about the definition? At least it shows that, in that counterfactual situation, the definition would be incorrect. And this is also what would happen with Tarski's definitions in those counterfactual situations imagined by Putnam where the words of the object-language gained different meanings: they would no longer define what is intended. Tarski explicitly recognizes this when he makes the defined predicate relative to a particular language with a fixed interpretation. To want the definition to keep its adequacy in such counterfactual situations would amount to want to freely change the reference of '*L*' in 'true in *L*', or requiring '*L*' to be a variable over languages. But Tarski never intended to define 'true in *L*' for variable *L*. Putnam's criticism, therefore, is justified only if we can say that he should have done that – even when we know that Tarski found it not possible,

[17] Davidson (1990) 288.

because in that case the metalanguage itself would be among the values of '*L*' and that would bring us back to the semantic paradoxes.

Obviously the talk of definitions as being correct or incorrect in this sense is possible only when definitions are not intended as stipulative. Etchemendy attributed the initial error of Davidson and his followers to 'the ease with which we read substantive content into what is intended as a stipulative definition'. In his view, 'definitions presuppose nothing', 'definitions make no claims, provide no information'[18]. But to say this is not only to completely miss Tarski's intention – who obviously didn't want his definition to be stipulative – but also to prevent oneself from understanding the way Tarski turned T-sentences into logical truths. For what Tarski did was exactly to include in the definition itself the empirical content, the semantic information, which is present in pre-theoretical T-sentences. And, there, the detour through satisfaction and the recursive technique were essential to get the packing of an infinite amount of information into a finite formula. Then, as in my snow example, it's because that empirical information has been assimilated by the definition itself that it disappears from T-sentences, these becoming mere trivialities. So Tarski's procedure seems to be just an instance of the usual process, described by Quine already in 'Truth by Convention', through which, as science advances, 'what was once regarded as a theory about the world [in this case: about the object-language] becomes reconstrued as a convention of language [in this case: of the metalanguage]'[19]. If we gave Tarski's definitions the status of conventions, they would have the effect of making T-sentences true by convention.

The problem is that not all empirical information is sufficiently secure to be included in a definition. In Quine's example, Einstein includes the information that the speed of light is constant in the definition of 'simultaneity at a distance'. But this piece of information has a very different status from that possessed by the information that snow is white or that 'Gavagai' means rabbit. Including whiteness in the definition of 'snow' would be as good science as to include heaviness and lightness in the definitions of 'earth' and 'fire'. The insufficiency we can point to Tarski's theory is, I think, that of having included in the definition of truth infor-

[18] Etchemendy (1988) 58.
[19] Quine (1976) 77.

mation of a kind that is so fragile and so badly understood that, as Quine saw, is itself the one that, first of all, desperately lacks explanation.

In my view, this explains in part the fact that Quine, in all his presentations of Tarski's theory, always chose situations in which the object-language is included in the metalanguage and, hence, in which Schema T (*i.e.* the schema that T-sentences instantiate) has a purely disquotational form, becoming true whenever the same sentence is written in its two blanks. For, in those cases, there is no place for any doubts concerning the interpretation of the language for which truth is being defined. Those are the cases where we can say that the predicate is 'transparent' and that truth is disquotation: the attribution of truth to the sentence 'Snow is white' is just as clear and intelligible as the attribution of whiteness to snow. When the information included in the definition has this purely disquotational character, the problems raised by Putnam and many others don't apply.

Disquotation, though, only works in very limited and uninteresting cases. In all other cases, when what we have in front of us are concrete speakers with more or less different languages or idiolects, the situation, rather than disquotational, is interpretative: to meaningfully attribute truth to a sentence written or uttered by someone, we need to know what it means for her[20]. Now, it's when we turn to this kind of situation (the radical scenario where interpretation begins at home and where translation, even when it is homophonic, is no less translation) that, on the hand, the insufficiencies of Tarski's theory become more evident and that, on the other hand, invoking it to defend a cleavage between the theory of reference and the theory of meaning reveals itself to be an unjustified move.

To show this, I should go back to Putnam's criticism and to the visible uneasiness with which he sees that his argument needs to use traditional notions and distinctions Quine has tried to undermine – such as the notion of what the extension of a Tarskian truth predicate would be in a possible world in which our sentences had different meanings, and even the distinction between a definition that is only extensionally correct and

[20] On the restricted scope of disquotation, see the discussion with Quine in Davidson (1997), minutes 37 to 47.

a definition that also captures the intuitive meaning of the defined term[21]. Aware of this problem, Putnam then attributes to Quine a reaction to his argument according to which 'If Tarski's notion [of truth] isn't the intuitive one, so much the worse for the intuitive one! [...] Tarski has given us a substitute for the intuitive notion that is adequate for our scientific purposes [...], and one that is defined in a precise way.'[22] Besides, one might add, in time scientific notions tend themselves to become more or less intuitive.

Here, I think that Putnam has understated his case, and that this way-out is not open to Quine. Because, if the goal is the construction of a scientifically adequate substitute for the vague intuitive notion of truth, Tarski failed it so much as Carnap failed the analogous goal for analyticity. In 'Two Dogmas of Empiricism', Quine considers a definition of 'analytic sentence' for a specific language L_0 having 'the form explicitly of a specification, by recursion or otherwise, of all the analytic [sentences] of L_0'[23]. His main criticism is that, with such an enumerative definition, we end up knowing which sentences is analyticity attributed to, but, if we didn't understand it before, we don't end up understanding, or understanding any better, what is it that the definition attributes to those sentences. This point looks so obvious that many people don't realize that, in making it, Quine is helping himself with precisely that distinction between analysis of a notion and description of its extension which Putnam also invokes with so evident bad-conscience. Besides, the same kind of criticism can also be applied to Tarski's definitions of truth, though with a small difference. As already remarked, Tarski's definitions don't say which sentences of the language are true. What they recursively specify is, for each sentence, what are the conditions in which truth should be attributed to it. But understanding the conditions in which truth is attributed to the

[21] See the remarks in Putnam (1994) 334 about 'the reaction of Quine'. At a certain point, Putnam even justifies his use of 'the traditional language of 'meaning,' 'conceptual analysis,' etc.' by claiming that 'Giving up the analytic/synthetic distinction isn't the same thing as giving up the distinction between a philosophical analysis of a notion and a description of its extension, or, at least, it isn't giving up *that* distinction altogether' (328n3). But compare this with what Quine (1980) 132 says about definition and definability.

[22] Putnam (1994) 334.

[23] Quine (1980) 33.

sentences is not enough to understand what is it that, given those conditions, we are attributing to them. Dummett said it thus: 'We cannot in general suppose that we give a proper account of a concept by describing those circumstances in which we do, and those in which we do not, make use of the relevant word, by describing the *usage* of that word; we must also give an account of the *point* of the concept, explain what we use the word *for*.'[24]

A possible reaction would be to claim that, in the case of truth, when we understand the conditions in which it is attributed to a sentence, there is nothing else to be understood. For to say that the sentence is true is just an indirect way of saying that those conditions hold. This amounts to adopting a deflationary conception, according to which truth is not a substantive notion. But if it were so, all this discussion would have no subject, because there would be no notion that Tarski should have had analysed or substituted. Though Quine's apparent sympathies for such a deflationary conception are a matter for controversy[25], I presuppose in my argument, as Putnam and Davidson also did (and as, according to them, also did Tarski), that truth is a genuine property.

To explain the notion of analyticity we would need, Quine says, a definition of 'analytic in *L*' for variable *L*. Carnap didn't provide it, and Tarski also didn't do it for truth. In spite of the similarity, Quine has claimed that the two cases are different because, while for truth there is such a paradigm as Schema T, which gives it a 'peculiar clarity', we have nothing analogous for analyticity[26]. But when Quine says this, he once more has in view Schema T *in its disquotational version*, and this, we already saw, drastically limits the scope and force of his defence. On the contrary, when we put ourselves in what I called an interpretative situation, we can see that the difference between the two cases is not that big or, at least, not big enough to justify the division of semantics in two incommunicable provinces.

[24] Dummett (1978) 3. And Dummett's next statement – 'Classifications do not exist in the void' – completely agrees with what Quine says about the purpose of specifying a certain class of sentences, be it a class of analytic sentences (p. 33) or a class of postulates (p. 35).

[25] See Davidson (1994).

[26] See Quine (1980) 138.

There are two possibilities: *either* we require, to be able to adequately define truth, an explanation in sufficiently general terms of the way in which the truth of sentences is also determined by their meanings, *or*, having no such explanation, we allow ourselves to fix the language and its interpretation, and to include that unexplained information in the definition itself of a truth predicate which, therefore, must stay relative to that language-and-interpretation. In the first case, what we are accepting is the task of dealing with one of the central concepts of the 'theory of meaning'. In the second case, we are presupposing that we know the meanings of all expressions and sentences of the object-language and that we are able to give, for each one of them, a synonymous expression or sentence of the metalanguage. But, if we are able to do this, we can also extensionally define synonymy as that relation which holds between those, and only those, expressions and sentences we gave as synonymous. And if we can do this for two different languages, we should also be able to do it inside one of them. Now, as Quine himself said, once we have defined synonymy, analyticity reduces, through it, to logical truth.

I will conclude by noticing once more that Tarski would not be very surprised with what I've just said. For, after all, it was him who stated that, if we had gone through the formalization of the metalanguage and of the metatheory, 'the exact specification of the meaning' of the term 'translation' which occurs in Convention T 'would present no great difficulties'[27]. Obviously we would not expect that specification to be elucidative of the notion – intuitive or scientific, it doesn't matter here – of translation. Neither should we expect that from his definition of truth.

Ricardo Santos
Instituto de Filosofia da Linguagem
Faculdade de Ciências Sociais e Humanas
Universidade Nova de Lisboa
Avenida de Berna, 26-C
1069-061 Lisboa, Portugal
rsantos.ifl@fcsh.unl.pt

[27] Tarski (1983) 188n1.

References

Davidson, D. (1984), *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford.

—— (1990), 'The Structure and Content of Truth', *Journal of Philosophy*, 87, pp. 279-328.

—— (1994), 'What is Quine's View of Truth?', *Inquiry*, 37, pp. 437-440.

—— (1997), *Donald Davidson in Conversation: The Quine Discussion*, The London School of Economics, London (videocassette).

Dummett, M. (1978), *Truth and Other Enigmas*, Duckworth, London.

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

Field, H. (2001), *Truth and the Absence of Fact*, Clarendon Press, Oxford.

Putnam, H. (1988), *Representation and Reality*, MIT Press, Cambridge (Mass.).

—— (1994), *Words and Life*, Harvard University Press, Cambridge (Mass.).

Quine, W. V. (1976), *The Ways of Paradox and Other Essays*, revised and enlarged edition, Harvard University Press, Cambridge (Mass.).

—— (1980), *From a Logical Point of View*, second edition, revised, Harvard University Press, Cambridge (Mass.).

—— (1985), *The Time of My Life*, MIT Press, Cambridge (Mass.).

—— (1992), *Pursuit of Truth*, revised edition, Harvard University Press, Cambridge (Mass.).

Soames, S. (1999), *Understanding Truth*, Oxford University Press, New York.

Tarski, A. (1944), 'The Semantic Conception of Truth', *Philosophy and Phenomenological Research*, 4, pp. 341-375.

—— (1983), *Logic, Semantics, Metamathematics*, second edition, Hackett, Indianapolis.