

## A TRUTH PREDICATE IN THE OBJECT LANGUAGE

William G. Lycan

University of North Carolina

The semantic paradoxes arise when the range of the quantifiers in the object language is too generous in certain ways. But it is not really clear how unfair to Urdu or to Hindi it would be to view the range of their quantifiers as insufficient to yield an explicit definition of ‘true-in-Urdu’ or ‘true-in-Hindi’. Or, to put the matter in another, if not more serious, way, there may in the nature of the case always be something we grasp in understanding the language of another (the concept of truth) that we cannot communicate to him. In any case, most of the problems of general philosophical interest arise within a fragment of the relevant natural language that may be conceived as containing very little set theory. Of course these comments do not meet the claim that natural languages are universal. But it seems to me this claim, now that we know such universality leads to paradox, is suspect.

–“Truth and Meaning”

Thus did Davidson reply to Tarski’s objection that a truth-theoretic semantics for a natural language will self-destruct when that language’s Liar sentence rolls down the input chute.

It is a curiously brief, compressed, and as Lepore and Ludwig say (2005, p. 133) perfunctory response to an apparently devastating problem. Obviously he was not daunted, much less devastated, though he did say he “wish[ed he] had...a serious answer” (1967,

p. 314). "...I will say only why I think we are justified in carrying on without having disinfected this particular source of conceptual anxiety."

### 1. LEPORE AND LUDWIG'S WAYS OUT

The foregoing famous passage contains at least two, probably three and possibly four different points, and it is hard to say which of its sentences contributes to which of those. My own breakdown would be linear: The first point is put in the opening two sentences; the third sentence ("Or,...") could be an attempt to restate that first point, but could also be taken as making a second one; the fourth sentence ("In any case,...") states a new point; and the concluding sentence ("But it seems to me...") suggests a further one.

Lepore and Ludwig unpack the passage into two overall appeals. The first is the one about quantifier ranges, which they gloss in this way: "[N]atural languages should not be regarded as having the expressive resources required to generate the paradoxes, despite appearances to the contrary" (p. 133). Second, whether not we take natural languages to be universal,<sup>1</sup> and even if they are subject to the paradoxes, "little is lost if we ignore the semantic vocabulary" (ibid.).

The point about quantifier range is this: A Tarskian truth definition quantifies over sentences of the object language, as in "For all sentences  $s$  of  $L$ , if ... $s$ ..., then  $s$  is  $T$  iff..." But "[i]f the quantifiers of the object language do not include within their scope [sic; meaning *range*] every sentence of the object language, particularly those with 'is  $T$ ', we could not generate the semantic paradoxes from a sentence of the above form stated in the language, nor could we define a truth predicate for the language as a whole" (pp. 133-34). Lepore and Ludwig speculate that Davidson had in mind a type theory built into the object

language, which type theory would enforce some version of Russell's Vicious Circle principle. "The suggestion, in a nutshell, then, is that natural language sentences are interpreted by their speakers so as to avoid vicious definitional circles" (p. 135). On this view, when we see a sentence metapredicating truth, such as "'Snow is white' is true" is true," we would have to disambiguate, and of course there would follow "an infinite hierarchy of interpretations of truth predicates, and correspondingly of quantifiers, which are understood differently depending on the restrictions on the values of their variables, imposed by what predicates they are used with" (ibid.).<sup>2</sup> But the paradoxes would not arise.

Lepore and Ludwig do not rest content with that, for "[i]t is at least tendentious to claim that speakers of natural languages have the sophistication this response to the problem presupposes. And it does not seem *impossible* for a natural language to be semantically defective in the way that would give rise to the semantic paradoxes" (pp. 135-36). (In fn 111 they add that of course the infinite hierarchy of interpretations suggests an antiDavidsonian failure of learnability as well.) So they move on to the point about little being lost.

Viz., they consider the move I had discussed in Lycan (1984, p. 25): Suppose we restrict our attention to the largest fragment of English that contains no semantical terms. That fragment will be nearly all of English, and will pose all the myriad semantical puzzles that linguists discuss. To give a truth-theoretic semantics for that fragment would be a monumental achievement, and it would have all the advantages that Davidson claimed for his sort of investigation—explaining semantical data, showing how the meanings of complex expressions depend on those of their parts, thereby showing how infinite capacity is generated from finite means, and so on. The worst one could say about the fragment

semantics is that it would not be complete; “our ambition to provide a meaning theory for natural languages would not be fully achieved” (p. 137).

Now, I myself do not agree that that is the worst that could be said of the incomplete theory. A truth theory for the largest fragment of English that contained no words for vegetables would be incomplete too, but in a way that no one would care about because the lack would have no theoretical significance. Moreover, obviously, the incomplete theory would not just happen to be incomplete; it would be incomplete on pain of contradiction were it to be extended. Our ambition to provide a meaning theory for natural languages would (so far as has been shown) still be impossible, not just not fully achieved.<sup>3</sup>

Notice that the advantages of the fragment approach would be preserved, and honor a bit better satisfied, if the whole language were treated, but in a paraconsistent logic whereby the Liar contradiction could simply be walled off. One’s overall theory would be inconsistent, but there would be no Explosion. (If one were a dialetheist, one could even insist that the theory is true.<sup>4</sup>)

In any case, happily, Lepore and Ludwig do not rest content with the fragment approach either. At this point (pp. 137-38) they advert to a claim they have defended earlier in their book (Ch. 4): that contra Davidson himself, it is not the truth theory for a language that gives the meanings of that language’s sentences; rather, the meaning theory for the language is derived from the truth theory by way of a formula that speaks of translations: “[1] For every sentence  $s$ , language  $L$ ,  $s$  in  $L$  means that  $p$  iff a canonical theorem of an interpretive truth theory for  $L$  uses a sentence that translates ‘ $p$ ’ on its right hand side” (p. 73, repeated 137). It now does not matter that the truth theory itself is formally inconsistent, Lepore and Ludwig maintain, because [1] can be used to extract “‘The Liar is not true in  $L$ ’

means that The Liar is not true in L”—just what we want—from “‘The Liar is not true in L’ is true in L iff The Liar is not true in L,” irrespective of the latter’s being a contradiction.

This is a neat solution, and in a way more Davidsonian than the ones Davidson himself had considered, especially because Davidson had himself required that the r.h.s. of a T-sentence be a “translation” of the mentioned target sentence. But I myself would rather continue to explore the possibility of fashioning a noncontradictory truth theory for a natural language that contains its own truth predicate.

## 2. MY PROPOSAL

One might think that any general solution to the Liar Paradox could be taken over into a Davidsonian semantics. And of course many theorists have made assaults on the Liar, usually enlisting heavy and powerful formal apparatus.<sup>5</sup> Needless to say I shall not even begin to survey those attempts, much less demonstrate the universal superiority of my own approach. I am concerned only to provide a “solution” to the Liar that is adequate to the facts of natural language, that is fairly simple, and that will make linguistic semantics safe from Tarski’s famous objection in the sense that it will keep Davidsonian truth definitions for particular natural languages from blowing up. (It is a linguist’s solution, not a logician’s. Nor is it intended to reveal anything conceptually deep.) My approach will start with the first attempt discussed by Lepore and Ludwig, and so will be hierarchical in the spirit of Tarski’s treatment of formal languages; but it will be shown to resist the usual objections to hierarchical approaches. *Contra* Tarski, there is a good and clear if somewhat grotesque sense in which a natural language can contain its own truth predicate without paradox.

As has been well known at least since the publication of Kripke (1975), there is no purely syntactic or semantic way of picking out sentences that generate Liar antinomies. Self-reference is not required for such antinomies, since pairs or triples of individually innocent sentences can contain Liar contradictions through cyclical inter-reference.<sup>6</sup> Moreover, whether a given set of sentences is Liar-paradoxical can depend on contingent, extralinguistic fact. Consider an example of a “contingent Liar cycle” in the sense of Kripke and also Barwise and Etchemendy (1987, p. 23):

$(\alpha_1)$  Max has the three of clubs.

$(\alpha_2)$   $(\beta)$  is true.

$(\beta)$  Either  $(\alpha_1)$  or  $(\alpha_2)$  is false.

The contradiction in this set would be no antinomy if  $(\alpha_1)$  were, as a matter of fact, false. But it is an antinomy because (we are assured)  $(\alpha_1)$  happens to be true.

Even contingent Liar cycles pose Tarski’s problem for natural-language truth definitions. For from the truth definition for any language containing  $(\alpha_2)$  and  $(\beta)$ , any contingent statement (in the present example, “Max does not have the three of clubs”) can be deduced—including the negation of some contingent statement that has already been deduced, at which point the truth definition blows up.

Tarski’s solution to the Liar for formal languages is syntactical, of course. He expunged “true” and “false” from the object language, kicking them upstairs to the metalanguage; the latter contains some English words, some logical symbols, quotation marks and a truth predicate.<sup>7</sup> (It also contains the entire object language, since we will need it to formulate all the instances of Tarski’s schema.) Thus for Tarski, when the Liar

sentence is formalized, it is seen to be ill-formed: Because it contains a translation of the word “false,” it is (by definition) a sentence of the metalanguage, not of the object language; but by stipulation, “true” and “false” apply only to sentences of the object language; it is simply ungrammatical to apply the same semantical predicates to a sentence of the metalanguage. Of course, sentences of the metalanguage *are* themselves true or false in the intuitive sense, but those semantic values can be assigned only in a *meta*-metalanguage, and so on up. Let us use “true<sub>0</sub>” as our original truth predicate, a word of the metalanguage that applies to sentences of the object language, and “true<sub>1</sub>” as our new word of the meta-metalanguage that applies to sentences of the metalanguage, and so on. Thus we generate the familiar potentially infinite hierarchy of successive metalanguages, each with its own truth predicate that applies to sentences of the next level down. In practice we never need more than one or two of these metalanguages, but we know they are there in principle, on pain of the Paradox.

Those of us who (unlike Tarski) are interested in natural languages may borrow his technique. If the Liar sentence is as ungrammatical in English as it is in Tarski’s formal languages, then our truth definition for English is excused from assigning it a truth condition; we need not allow it into the blank in Tarski’s schema in the first place. And if it is grammatical but only by dint of some special semantic stage-setting, its resulting semantics may succeed in removing the contradiction from our truth theory. Let us then distinguish within English between the nonsemantical part, which has ordinary words but no metalinguistic expressions like “true,” and a metalinguistic part which includes the latter expressions. English, we might say, is really an amalgam of two separate languages, and we are skilled at speaking both of them intermixed, just as some residents of the American

Southwest speak an amalgam called “Spanglish.” Of course (again), some of the metalinguistic sentences we can form in English can rightly be called *true* also, so we are off and running with “true<sub>1</sub>,” “true<sub>2</sub>” and all the rest.<sup>8</sup>

But that just brings us to the obvious objection: My hierarchical suggestion has a theoretical liability, that of being *false*. English just is not like that. For a formal language, we can stipulate that such-and-such an expression will not be counted as well-formed. But for a natural language we can stipulate nothing at all; it was here first, and our job was just to learn it (and perhaps to investigate it). Having learnt it, we find that English in particular is *not* an amalgam of infinitely many different metalanguages, each having its own truth predicate. When we see the word “true” in an English sentence we do not have to worry about which of the infinitely many metalanguages it is written in. “True” just means *true*.<sup>9</sup>

Let us expand this last point. According to my current proposal, English (despite its appearance of unity) is subdivided into infinitely many sublanguages, each with its own truth predicate. These truth predicates cannot be substituted for each other, on pain of ungrammaticality, and so they cannot be regarded as mutually synonymous. We have to understand them as being homonyms, different words that happen to be spelled the same way, like “die,” “die” and “die.” But homonyms have to be learned separately. So, if the English expression “true” is really homonymous as between “true<sub>0</sub>,” “true<sub>1</sub>” and the rest, each of its uses has to be learned separately. And this is impossible, since there are infinitely many such uses. Our proposed solution fails.<sup>10</sup>

However, we need not, and I do not, grant that the various truth predicates are *sheerly* ambiguous, like “die,” “die” and “die.” They are at worst paronymous, since their meanings are, to say the least, closely analogous, and clearly one could learn “true<sub>n</sub>” by

analogical stretching from “true<sub>n-1</sub>.”<sup>11</sup> But in what follows I shall be able to finesse this issue.

Recall that in any case “true” is relative to a language; a sentence is not just true or false, period, but true or false in such-and-such a language. Therefore we need not regard “true” as being *homonymous as between* all the different metalanguages; we can simply *relativize it to* a metalevel, as indeed we have implicitly had to do all along. Thus, let us replace “true<sub>0</sub>” by “true-in-*O*,” where *O* is the object-language part of English, and replace “true<sub>1</sub>” by “true-in-*M*<sub>1</sub>,” where *M*<sub>1</sub> is the first metalanguage, and so on. (Think of these new expressions exactly on the model of “true-in-French,” “true-in-German,” etc.) This recaptures the feeling that “true” in English just means *true*—it is the same word everywhere, univocal and not a homonym. On the other hand, we are treating “true” as a relative term; strictly speaking, if someone says that some sentence is true, we have to ask, “True in what language?” But remember, this has *always* been the case, and we have managed to live with it up till now.

Let (L) be “(L) is false.” (L) now suffers death by disambiguation. “(L) is false” is not a sentence of *O*, for *O*’s lexicon contains no semantical terms such as “false,” not even language-relativized ones. Thus, as Tarski intended, the string (L) is simply ill-formed if considered as written in *O*.

So far as it is well-formed, (L) purports to be a sentence of the metalanguage *M*<sub>1</sub>, in which case it means “(L<sub>≈1</sub>) is false-in-*O*” (where “(L<sub>≈1</sub>)” means “(L) considered as a surface structure in *M*<sub>1</sub>”). But then it is false(-in-*M*<sub>1</sub>), since it is not a sentence of *O* at all. But neither could “false” in (L) itself usefully mean “false-in-*M*<sub>1</sub>,” because no sentence of *M*<sub>1</sub> mentions *M*<sub>1</sub>; (L) would then have to be taken as a sentence of *M*<sub>2</sub>, in which case (as

before) it would come out false(-in- $M_2$ ) because it would entail of itself that it was a sentence of  $M_1$ . (*Und so weiter.* (L) could of course be disambiguated as a sentence of a higher-levelled metalanguage than  $M_1$ , say  $M_6$ ; in that case it would mean “(L $\approx$ 6) is false-in- $M_5$ ,” and would still be false(-in- $M_6$ ) because it is not a sentence of  $M_5$ .)<sup>12</sup>

And the Tarski biconditional “‘(L) is false-in- $O$ ’ is true iff (L) is false-in- $O$ ” is a sentence of  $M_2$ , since it is directed upon a target sentence of  $M_1$ , and its truth predicate means “true-in- $M_1$ ,” in which case the biconditional is not contradictory at all but perfectly true; “(L) is false-in- $O$ ” is true-in- $M_1$  iff (L) is false-in- $O$ . (Both sides are false[-in- $M_2$ ].)

The Strengthened Liar, (L’) “(L’) is not true,” goes much the same way. It is not a sentence of  $O$ , for it contains a semantical term that is no lexeme of  $O$ .<sup>13</sup> Prima facie it is a sentence of  $M_1$ , in which case it means “(L’ $\approx$ 1) is not true-in- $O$ ”; but then it is true(-in- $M_1$ ), since it is not a sentence of  $O$  at all. Nor could “true” in (L’) itself usefully mean “true-in- $M_1$ ,” because no sentence of  $M_1$  mentions  $M_1$ ; (L’) would then have to be taken as a sentence of  $M_2$ , in which case (as before) it would come out false(-in- $M_2$ ) because it would entail of itself that it was a sentence of  $M_1$ —*und so weiter*.<sup>14</sup>

### 3. THREE BAD OBJECTIONS

First objection: My title promised a truth predicate in the object language. Piquant advertising, since everyone thinks Tarski simply proved by *reductio* that there cannot be a truth predicate in the object language. Piquant and *false* advertising, since Tarski did prove that there cannot be a truth predicate in the object language. Moreover, as a matter of public record, I have myself accepted his hierarchical alternative format for semantical predication and (precisely) kicked semantical terms out of the object language.

Well, yes. Of course there cannot be a truth predicate in the object language strictly so called. But the natural language we call *English* can contain “true” as an English word. The individuation of “languages” is a notoriously slippery matter (cf. Lycan [1984, pp. 70-71]), and in talking of languages we do often use coarse-grained counting policies as well as fine-grained ones.

For example, we say of Roger, an acquaintance of mine, that his *only* language is English (he knows no Spanish, German, or other foreign tongue); but in fact Roger’s written English and his spoken conversational English differ both lexically and grammatically, and he is also fluent in urban Black dialect when he chooses to use it. Does he then speak one language, or three? It simply depends on how we choose to count. Because his (at least) three dialects are so closely related to each other and so distant from any Spanish, German etc. dialects, it is convenient for most everyday purposes to count them *as* dialects of a single language and call that language just “English.” But for purposes of linguistic theory, since the three dialects differ in their syntax and in their vocabulary, we treat them separately and each gets its own grammar and truth theory, even though there is lots of overlap.

So too, we may treat  $M_1$  as distinct from  $O$ ,  $M_2$  as distinct from both, and so on up. Their grammars and their truth theories will overlap heavily, just as Roger’s dialects do albeit more so, but for theoretical purposes they are numerically distinct. At the same time, for everyday purposes we lump them together and call the speaker-hearer’s language “English.” There is nothing illicit, sloppy or fictional about that practice; it is simply coarse-grained individuation. And it is in that sense that, on my view, English contains its own truth predicate.

Second objection: But then, which of my denumerably many nested languages *is English*? Not *O*, since “true” and “false” are English words, but not  $M_1$  to the exclusion of  $M_2$  or vice versa; and since  $M_1$ ,  $M_2$  and the rest are all distinct languages, it *follows* that at most one can be English, but none has any better claim than any other.

Whoever raises this point was not listening to my reply to the first objection. Of course the term “English” as it is used in English does not designate any one of the hierarchically nested Tarskian languages I have posited, any more than it specifically designates Black dialect or BBC English or Down East or West Indian. Quite independently of my theory, “English” *always* only vaguely indicates a melange of dialects that are theoretically distinct. The second objector’s question makes no more sense than would that of which English dialect “is English.”

Third objection: My theory is psychologically impossible, or at least outrageous. I have tried to transport Tarski’s infinite hierarchy of metalanguages to the practice of natural language semantics; but then according to my own view of linguistic semantics as part of a theory of speakers and hearers and hence as psychologically constrained (Lycan [1984, Chs. 9 and 10]), there would have to be an infinite hierarchy of metalanguages in some sense impacted within every normal human speaker-hearer’s brain, and a discriminable linguistic competence to match each one. Subjects would have to have internal representations of the metalanguages  $M_1$ ,  $M_2$ ,... to infinity, and flag every semantical term with a particular one of those representations every time the term was tokened externally or internally. And that does not sound like anything that actually happens within the human organism; it might not even be metaphysically possible for it to happen within a brain.

Fortunately, for my purposes I do not need an actual infinite Tarskian hierarchy in the mind. (This is one way in which my “solution” is for linguists and psychologists, not for logicians.) For there is an historical limit to the level of metalinguistic predication ever actually performed by a human speaker, and no doubt there is a psychological and/or psychobiological limit to the level that could be performed even by a very gifted speaker. Linguists, philosophers and mathematicians may occasionally have used  $M_6$  to discuss  $M_5$ , but that is very unusual behavior, and it is unlikely that anyone has ever used  $M_{37}$  to discuss  $M_{36}$ ; it is unlikely that anyone save perhaps Lewis Carroll or Nuel Belnap has ever mentally represented  $M_5$ .<sup>15</sup> The objector is right to assume that any metalanguage that is represented is represented in the brain, but few metalanguages have been represented at all.

My view does imply that everyone has an internal general concept of the relation of metalanguage to object language, and so implicitly the idea of an  $M_{37}$  directed upon  $M_{36}$ ; the hierarchy must be *potentially* realized in normal speaker-hearers. I think that consequence is true, since the required notions of implicitness and potentiality are very weak, entailing nothing about the actual representation of many higher-level metalanguages.

#### 4. A TOUGHER OBJECTION

The previous three objections misunderstood the structure of my proposal. But here is a fourth, which does not:<sup>16</sup> Sometimes we predicate truth of a conjunction whose conjuncts are of different levels, or of a sentence-token whose containing language we do not know, or of a bound variable that may range over sentences from different languages.

E.g.:

(1) “ ‘Snow is white’ is true, and grass is green” is true.

(2) What Mary said is true,

asserted because we are confident of Mary’s sincerity and reliability on the topic, without knowing what it was she said or what language she said it in.

(3) Everything John said tonight was true,

where John made a number of assertions, most in the object language but some of them themselves truth predications. In such a case we may have excellent grounds for our own truth predication but no notion at all of which level of my Tarskian hierarchy we are discussing. Thus there is no nonarbitrary way of determining the appropriate level.

This is a nasty problem for some hierarchical solutions as they would apply to natural languages, viz., for any solution that posited successive truth predicates, “true<sub>0</sub>,” “true<sub>1</sub>,”..., as distinct and primitive (though paronymous) morphemes from distinct languages. We would have to find a principled way of deciding which one of those morphemes was expressed by “true” as it occurs in (1), then which one is expressed in (2), then which in (3), and so on for every other such example. Also, and worse, it does not seem that any of the successive truth morphemes could univocally be expressed by the concluding “true” in (1) or in (3), for that “true” crosses levels and none of the morphemes is allowed to do that.

The problem is not nearly so bad for my own theory, since I treat “true” itself as univocal and locate my hierarchy rather in the subscripted names of my nested languages. Moreover, since my “true” is relative and must always be filled in by an overt or covert language parameter, we have some options available. “True” can be filled in by the name of

a specific language or by a bound variable, so we could try various names, and variables bound in various ways.

Names will not work, for we often do not know which language we are discussing. In the case of (2), for example, we do not know which language Mary used. We may even have heard her utterance (and failed to understand it) without knowing what language she was speaking. Thus we are not in a position to interpret (2) as anything of the form, “What Mary said is true-in-L,” where “L” is replaced by the proper name of a language. Variables it will have to be.

A first suggestion would posit simple existential quantification; “true” would be understood as meaning “true in some language or other.” Our three test cases would be interpreted as follows:

$$(1E) \quad (\exists \square L)T[“ ‘Snow is white’ is true, and grass is green”, L]$$

Or, regimenting also within the mention quotes,

$$(1E') \quad (\exists L)T[“(\exists L^*)T[‘Snow is white’, L^*] \& \text{Grass is green}”, L]$$

$$(2E) \quad (\exists L)T[(\iota S)(\text{Said}(\text{Mary}, S)), L]$$

(3), existentially read, has a scope ambiguity:

$$(3E_{wk}) \quad (S)(\exists L)(\text{Said}_{(\text{tonight})}(\text{John}, S) \supset T[S, L])$$

$$(3E_{str}) \quad (\exists L)(S)(\text{Said}_{(\text{tonight})}(\text{John}, S) \supset T[S, L])$$

Let us see whether the existential proposal works.

If “language” is taken very broadly in this discussion, as including every formal structure of <lexicon, semantic interpretation, grammar> that exists in Plato’s Heaven—merely possible languages, if you like—then (1E)-(3E) are obviously too weak, for each is trivially true. Rather, “language” must comprehend only tongues that are actually spoken and understood by sentient beings.

Even so, consider (2). (2E) might be made true by the coincidental occurrence of Mary’s sentence in an actual language other than the one she was using, even though her sentence was false in the language she was using and hence (2) is intuitively false. So (2E) as an analysis is uncomfortably hostage to the facts, viz., to the nonexistence of an alternative containing language in which Mary’s sentence occurs and is true. We might think the latter fact to be a very robust one; Davidson’s (1969, p. 163) example, “Empedocles leaped” (English) / “Empedokles liebt” (German), is contrived at best. But if my account of metalanguages as dialects is correct, one and the same sentence occurs in each of several distinct languages within the same person and also across different speakers of the same natural language. (Of course normally, when it does so, it keeps the same meaning, so we would not have a potential case of differential truth-value. But the truth predicate itself could supply a difference; my main thesis is that “S is true” means different things, and can change its truth-value, depending on which metalanguage it is taken to inhabit. E.g., “ ‘ “Snow is white” is true’ is true” is itself true if it is disambiguated as “ ‘ “Snow is white” is true[-in-*O*]’ is true-in- $M_1$ ,” a sentence of  $M_2$ , but is false if understood naïvely as “ ‘ “Snow is white” is true[-in-*O*]’ is true-in-*O*,” a sentence of  $M_1$ . If what Mary tokened was that sentence and she meant it in the second, preTarskian naïve sense, (2) should come out false, but (2E) would be true.)

(1) itself is true as it stands, so (1E) should come out true, and it does; it is verified by the language  $M_1$ . But we know any problem would stem, contrariwise, from the logical weakness of (1E): Could some (1)-analogue be false though its existential analysans were true? Here again we run into Davidson's problem of cross-linguistic soundalikes. Suppose contrary to legend that Empedocles did *not* leap from Etna, but that he is a lover (in the afterlife, if you are bothered by tense). Then “ ‘Empedocles leaped’ is true[-in- $O$ ] and grass is green” is false, but its existential analysans is true because “Empedokles liebt” is true-in-German. (If you are again unmoved by Davidson's own contrived example, still thank me for sparing you the less artificial and more convincing but hideously convolute substitute involving nested metalanguages as in “ ‘ ‘Snow is white’ is true’ is true.”)

Turning to (3), suppose John said each of the following: “Grass is green,” “ ‘Snow is white’ is true,” and “ ‘ ‘Roses are red’ is true’ is true.” Then again both (3Ewk) and (3Estr) will be true whenever (3) is, for (3Estr) is verified by the language  $M_2$ : “Grass is green,” “ ‘Snow is white’ is true[-in- $O$ ],” and “ ‘ ‘Roses are red’ is true[-in- $O$ ]’ is true[-in- $M_1]$ ” are all true-in- $M_2$ , which contains all of  $O$  plus the truth predicate and the names “ $O$ ” and “ $M_1$ .” But just as with (2), even (3Estr) is insufficient for (3) because of Davidsonian soundalike cases (imagine that one of the things John said, but wrongly, was “Empedocles leaped”).

Thus, the existential readings do not help. But moving on to universal quantification does not help either. The universal method offers the following paraphrases.

(1U) (L)T[“ ‘Snow is white’ is true, and grass is green”, L]

Or, again regimenting within the mention quotes,

(1U') (L)T["(L\*)T['Snow is white', L] & Grass is green", L]

(2U) (L)T[( $\iota$ S)(Said(Mary, S)), L]

(3U) (S)(L)(Said<sub>(tonight)</sub>(John, S)  $\supset$  T[S, L])

(1U) and (1U') are immediately counterexamples, for they would not be sentences of *O* and therefore would fail to be true-in-*O*. If what Mary said was a sentence of the object language, (2U) is materially equivalent to (2); but if what she issued was instead a truth predication, (2U) is counterexampled in the same way (1U) was, and (3U) is a nonstarter for the same reason.

Instead let us try *proprietary* quantification, and bind our language variable with a definite description operator, expressed by the Russellian iota. The idea will be that a given sentence-token will be true-in *its* own containing language. This of course requires the assumption that every sentence-token has, associated with it, a single designated or proprietary language. But that assumption is not implausible, for we may suppose that each sentence-token was produced by its utterer in a particular language.<sup>17</sup> Thus, for any such token we can speak of *the* language in which it was produced, ( $\iota$ L)Prod(S,L).

(1P) T["'Snow is white' is true, and grass is green", ( $\iota$ L)Prod(" 'Snow is white' is true, and grass is green", L)]

Or

(1P') T["T['Snow is white', ( $\iota$ L)Prod('Snow is white', L)] & Grass is green", ( $\iota$ L)Prod("T['Snow is white', ( $\iota$ L)Prod('Snow is white', L)] & Grass is green", L)]

(2P)  $T[(\iota S)(\text{Said}(\text{Mary}, S)), (\iota L)\text{Prod}(S,L)]$

(3P)  $(S)(\text{Said}_{(\text{tonight})}(\text{John}, S) \supset T[S, (\iota L)\text{Prod}(S,L)])$

Will these do? Mechanically, it seems, yes. (1) says that “ ‘Snow is white’ is true *in its own designated L* and grass is green” is true in *its* designated L, which seems right, assuming “Snow is white” is true-in-*O* and the whole quoted containing sentence is true-in- $M_1$ . (2) is now understood as saying that what Mary said was true in its designated L, whichever language that was; thus we do not need to know the level of Mary’s token. (3) says that everything John said was true in its respective L, whatever jumble of levels might be represented by John’s collected discourse. So far, at least, our current proposal is adequate to the data.

But it is not perfect.

## 5. TWO FURTHER DIFFICULTIES

Two problems stand out, the most obvious of which results from the fine grain of my hierarchy of metalanguages. My “designated language” assumption is not merely the claim that someone’s token, say of /ɛmpədɔkliyz liypt/, was issued in English rather than in German or vice versa. For the assumption to serve its purpose it must be understood as meaning that every token is produced at a particular level of its utterer’s Tarskian hierarchy. And given the (face it) *ad-hocness* of my hierarchical hypothesis for natural languages, one may suspect that for many a token at least, any choice of a designated level for that token would be arbitrary; thus there is not, in the real world, a single proprietary “language” (in

the finely individuated sense) for every token. And therefore I am unentitled to my descriptors containing “Prod.”

But in the present context we need not take charges of arbitrariness and *ad-hocness* too seriously. For one thing, we are dealing with a bonafide paradox or antinomy. In addition, I am not promising conceptual progress or philosophical enlightenment, but only trying to keep truth definitions from blowing up. If I can merely *formulate* a not unreasonable policy for level-assignment and motivate it a bit, I can establish a genuine mapping from sentence-tokens into the language levels, and the “Prod” relation will be a real one, whether or not anyone supposes it exists in the brain independently of our conceiving it.

And here is a simple policy: Assign each sentence to the lowest level possible. This means that sentences containing no semantical terms will automatically count as produced in  $O$ . Sentences containing single semantical predications directed upon expressions of  $O$  will automatically be assigned to  $M_1$ . (That will hold even if such sentences have conjuncts that alone contain no semantical terms; recall that higher metalanguages all contain  $O$  gratis.) Embedded semantic predications such as (1) will be assigned to as high a level as is needed to respect our hierarchical type restrictions, but no higher—thus (1) will be classified as a sentence of  $M_2$ . Sentences which quantify nonspecifically over other utterances will be assigned to as high a level as is needed to accommodate the various type restrictions induced by the actual occurrences of semantical terms in those other utterances, but no higher.

That is my assignment policy formulated. Now to motivate it, after a fashion. Right: (i) It works. (I think.) That is, it should keep truth theories from blowing up. (ii) It

is simple, being stated in only eight words. (iii) It is natural, in that the object language has a natural priority in the real world. Semantic predications are in an obvious way parasitic on the object language; also, the object language and  $M_1$  are concretely real, while the infinitely many distinct higher metalanguages are less so, being artifacts or constructs of a stilted Tarskian approach to natural language. (iv) The policy is psychologically plausible. As is granted all around, few of my hierarchical distinctions are realized in the brain. For the most part we speak and think in  $O$ . Sometimes we ascend to  $M_1$ , but very rarely higher, and probably we are incapable of ascending much higher than  $M_5$  or  $M_6$  in the course of ordinary language processing. (We can conceive, for theoretical purposes, of  $M_{327}$  and equally  $M_{110,679}$ , but that is psychologically a different matter from whatever mental reference to higher metalanguages might occur in ordinary people's everyday thinking and understanding.) Moreover, I should think, semantic ascent costs some psychological effort, and is to be avoided when unnecessary.

There. That should be enough to motivate my assignment policy, and the policy seems enough to support our “designated language” assumption. I do not anticipate any difficulty in formalizing the policy.

But now we must face the second problem of the two I mentioned. It is that I have played fast and loose with the type/token distinction. I have couched my response to the “tough objection,” and my new assignment policy, in terms of sentence-tokens rather than sentence-types. That was because our problem examples (2) and (3) seemed to concern tokens, or at least utterances made on specific occasions, and also because focusing on tokens is a good way of finessing difficulties created by ambiguity and deixis. But (1) forces the issue: The sentence it mentions need not have been tokened, recently or ever; I

doubt that that sentence has in fact ever been tokened in its own right. And even if that mentioned sentence has been tokened by someone at some point, no particular token of *its* mentioned sentence “Snow is white” is at hand, so there is no token there to fix a “language” in which the sentence is to be true. In general, a sentence evaluated as true or false need not have been “produced” on the occasion of evaluation, or even tokened ever in the history of the universe prior to its appearance between current mention-quotes. There need not have been any language in which such-and-such a sentence “was produced,” because the sentence may not have been produced at all.

Rather, the sentence mentioned in (1) and the shorter sentence mentioned in it are evidently being considered as types. That is legitimate; in natural languages, truth-values are sometimes predicated of sentence-types as well as of tokens.<sup>18</sup> Any solution to the Liar Paradox must accommodate that fact. We must extend our apparatus to cover types.

One way of doing that would be the backward way, viz., to reformatize sentences such as (1) into paraphrases about tokens; we might replace “‘Snow is white’ is true” in (1) by “Every token of ‘Snow is white’ would be true,” or the like. But we need not so complicate our paraphrases. Our present problem is really just to preserve the “designated language” assumption for types as well as for tokens, in the face of the fact that most types are never produced by actual speakers; and we can accomplish that task nearly by brute force.

Every natural-language sentence is a sentence of some natural language. When you first read (1), there was no doubt in your mind that “Snow is white” was a sentence of English in particular. Thus the “designated language” assumption is already looking good,

for the case of sentence-types. Of course, there are our two difficulties from earlier on: the problem of soundalikes, and the problem of fine-grained metalanguages.

Suppose there were a natural language quite different from English, an OSV language that happened to contain the lexemes “snow,” meaning *kitty litter*, “is,” a plural noun meaning *verificationists*, and a verb “white,” meaning *to eat*.<sup>19</sup> Then (1) would be ambiguous, and its new distinct reading would be (literally, so far as I know) false. But notice that the ambiguity appears *within English*, since (1) itself would be an English sentence on either reading. Such an ambiguity poses no threat to the “designated language” assumption.

What of the sentence “Snow is white” itself? It is now ambiguous as between its English meaning and its OSV meaning; that ambiguity is cross-linguistic and so does require qualification of the assumption. The assumption must be relativized to a contextual disambiguating choice. Such choices are made constantly, perhaps invariably, in language understanding, independently of the present issue. Even when we lack access to a speaker’s intentions, we hypothetically disambiguate and discuss either reading or both. Such is our way with ambiguity. Which way we would disambiguate “Snow is white” for purposes of Tarskian truth theory would depend on the natural language for which we were bent on providing a truth definition, in this case English rather than the OSV language.

If anything imperils the “designated language” assumption for sentence-types, it is the fine-grained horde of metalanguages, construed as “dialects of” a given natural language. Even if “Snow is white” is being remorselessly taken (in the context) as the English sentence rather than the OSV sentence, it has still not been identified as between  $O$ ,  $M_1$ ,  $M_2$ ,  $M_{879}$  and the rest. It is and will remain a grammatical sentence of every language in

the hierarchy. The “Prod” relation as originally intended will not help us assign “Snow is white” a designated language, for it does not apply to types and we have left actual tokenings behind. How, then, to support the “designated language” assumption?

I suggest we simply stick by the simple assignment policy I motivated for sentence-tokens (“Assign each sentence to the lowest level possible”). Items (i)-(iii) of the motivation apply to types as well as tokens. Item (iv), psychological plausibility, does not carry over directly, but it does so indirectly: If the type “Snow is white” *were* tokened, then other things being equal it would be produced in  $O$  rather than a metalanguage.

Let us mark the expansion of our designated-language apparatus by substituting “Iden” for “Prod”; “Iden(S,L)” is to be read as “S is identified as a sentence of L.” Thus “S is true” will always be expanded as “T(S, (ιL)Iden(S, L)).”

According to our conservative assignment policy, sentence (L) would be identified as a sentence of  $M_1$ , for it is ill-formed in  $O$  but need not be raised to any level higher than  $M_1$ . Thus, as was explained in section 2, it is false(-in- $M_1$ ).

## 5. CYCLES

Our truth-definition-saving technique is now complete. Let us apply it to Liar cycles, beginning with the simplest:

( $\delta$ ) ( $\epsilon$ ) is false.

( $\epsilon$ ) ( $\delta$ ) is true.

( $\delta$ ) contains a semantical term, so it is not a sentence of  $O$ . But it need not be raised any higher than  $M_1$ , and so by our conservative assignment policy, we locate it at  $M_1$  forthwith.

Proceeding on to  $(\varepsilon)$ , we see that  $(\varepsilon)$  predicates falsity of  $(\delta)$ , which is a sentence of  $M_1$ , and so we must identify  $(\varepsilon)$  as a sentence of (at least, and again at most)  $M_2$ .

Thus our cycle will be spelled out as

$(\delta')$   $(\varepsilon')$  is false-in- $O$ .

$(\varepsilon')$   $(\delta')$  is true-in- $M_1$ .

And  $(\delta')$  and  $(\varepsilon')$  pose no threat to the truth definitions for  $O$ ,  $M_1$  and  $M_2$ . Taken together, the truth definitions do entail the falsity of “ $(\varepsilon')$  is false-in- $O$ ,”<sup>20</sup> but that is fine, for  $(\varepsilon')$  is not a sentence of  $O$  and therefore, trivially, is not false-in- $O$ . The truth definitions remain jointly compatible.

The asymmetry between  $(\delta)$  and  $(\varepsilon)$ — $(\delta)$  being identified as a sentence of  $M_1$  and  $(\varepsilon)$  being kicked up a second flight to  $M_2$ —is, in an obvious way, gratuitous:  $(\delta)$  only happened to be treated first, because it happened to be listed first. Had we begun with  $(\varepsilon)$ , we would have located it in  $M_1$  and then booted  $(\delta)$  up to  $M_2$ .

But that is fine, for a parallel solution would then have presented itself. We would have

$(\varepsilon'')$   $(\delta'')$  is true-in- $O$ .

$(\delta'')$   $(\varepsilon'')$  is false-in- $M_1$ .

As before, the truth definitions would be seen to entail a trivial truth, “Not:  $(\delta'')$  is true-in- $O$ ,” and there would be no contradiction.

Our “designated language” assumption is slightly embarrassed again, for we end up assigning different languages to each of the component sentences, depending on the order in which we consider the Liar cycle’s members. This shows that my conservative assignment policy does not tamp down assignments as far as we would like, and there is still some arbitrariness in the designations of our designated languages. But the alternative choices disjoin nicely: It does not matter which order we take; either choice dispels paradox. In real life, the sentences ( $\delta$ ) and ( $\varepsilon$ ) would be tokened in one order or the other (else neither would designate the other), and that would determine their exact semantic interpretation.

Turning back to our contingent cycle:

( $\alpha_1$ ) Max has the three of clubs.

( $\alpha_2$ ) ( $\beta$ ) is true.

( $\beta$ ) Either ( $\alpha_1$ ) or ( $\alpha_2$ ) is false.

The first fact to grasp is that ( $\alpha_1$ ), identified as a sentence of  $O$ , is true, i.e., true-in- $O$ . Let us consider ( $\beta$ ) next, for ( $\beta$ ) directly mentions ( $\alpha_1$ ); thus we shall identify ( $\beta$ ) as a sentence of  $M_1$  and take “false” in ( $\beta$ ) to mean “false-in- $O$ .” ( $\alpha_2$ ) then will be a sentence of  $M_2$ , since it predicates truth of ( $\beta$ ). The cycle will thus be understood as:

( $\alpha_1$ ) Max has the three of clubs.

( $\alpha_2'$ ) ( $\beta'$ ) is true-in- $M_1$ .

( $\beta'$ ) Either ( $\alpha_1$ ) is false-in- $O$  or ( $\alpha_2'$ ) is false-in- $O$ .

But  $(\alpha_2')$  and  $(\beta')$  are both false, since  $(\alpha_1)$  is true-in- $O$  and  $(\alpha_2')$  is not a sentence of  $O$ .

And the truth definitions for  $O$ ,  $M_1$  and  $M_2$  survive, since the contradiction one would have derived, “ $(\alpha_2')$  is true iff  $(\alpha_2')$  is false,” is disambiguated as “ $(\alpha_2')$  is true-in- $M_2$  iff  $(\alpha_2')$  is false-in- $O$ ,” which is a truth.

What if we had taken  $(\alpha_2')$  and  $(\beta')$  in the opposite order? If, unnaturally, we refrain from looking ahead to what  $(\beta')$  says, we can identify  $(\alpha_2')$  as a sentence of  $M_1$ . Then since  $(\beta')$  applies “false” to  $(\alpha_2')$ , we must locate  $(\beta')$  in  $M_2$ . The paradox would then be explicated as:

$(\alpha_1)$  Max has the three of clubs.

$(\alpha_2'')$   $(\beta'')$  is true-in- $O$ .

$(\beta'')$  Either  $(\alpha_1)$  is false-in- $M_1$  or  $(\alpha_2'')$  is false-in- $M_1$ .

Since  $(\beta'')$  is not a sentence of  $O$ ,  $(\alpha_2'')$  is false(-in- $M_1$ ).  $(\beta'')$  is entailed by that and so is true(-in- $M_2$ ) but redundant. And, once again, the proof that is supposed to derive the contradiction from the truth definitions fails.<sup>21</sup>

Finally, let us address the nastiest cycle I have encountered, Kripke's (1975, pp. 695-97) Nixon/Dean example. John Dean asserts

(D) All of Nixon's utterances about Watergate are false.

And Nixon asserts

(N) Everything Dean says about Watergate is false.

According to Kripke's exposition, "Dean, in asserting the sweeping [(D)], wishes to include Nixon's assertion [(N)] within its scope (as one of the Nixonian assertions about Watergate which is [*sic*] said to be false); and Nixon, in asserting [(N)], wishes to do the same with Dean's [(D)]." Moreover, in real life, we might assess (D) as true and (N) as false or vice versa, depending on our empirical views. Yet we know nothing of levels (perhaps one thing Nixon had said was, "Haldeman told the truth when he said that Dean lied").

Now, it seems to matter a good deal in what order the relevant tokens were produced and with what intention. It is hard to imagine a realistic situation in which, simultaneously, Dean includes a token of (N) within the range of a token of (D) and Nixon includes the relevant token of (D) within the range of the relevant token of (N). The best I can do is to read both sentences timelessly, as ranging over all tokens Nixon or Dean have produced or ever will produce concerning Watergate.

Even so, those tokens will have come and continue to come in a temporal order, and our conservative assignment policy will apply. Past and present utterances pose no new problem. Whichever of (N) and (D) is uttered first, the token will be identified as a sentence of the lowest-level metalanguage as is consistent with accommodating all the semantical predications produced up till that point, and the other member of the pair would then be kicked one level up.

The *prima facie* trouble is with future utterances. For we simply cannot predict the highest metalevel at which Nixon or Dean might speak later on. (For all we knew at the time, in old age Nixon might have developed an interest in Lewis Carroll and/or fixed point theorems, and occasionally free-associated about Watergate while conversing in a very high-levelled metalanguage.) As a rock-ribbed realist about the future, I assume there is a

fact of the matter concerning that highest metalevel to come, but obviously there is not *at present* a metalevel of semantic predications occurring in Dean's or Nixon's brain that is geared to future semantical utterances.

Thus (granted), (D) and (N) make no specific reference to any future metalevel. But remember our strategy of proprietary quantification. (D) means

$$(DP) \quad (S)(\text{Asserts}(\text{Nixon}, S) \ \& \ \text{About}(S, \text{Watergate}) \supset F[S, (tL)\text{Iden}(S, L)]).$$

And (N) means

$$(NP) \quad (S)(\text{Says}(\text{Dean}, S) \ \& \ \text{About}(S, \text{Watergate}) \supset F[S, (tL)\text{Iden}(S, L)]).$$

Again, the semantical predications are relativized to the designated languages in which all the lower level assertions were *or will be* produced. Nixon, Dean or their brains can refer in that nonspecific way to the designated metalevels of future utterances without knowing what metalevels those will be.

I do here as before rely on the “designated language” assumption, and the example does still create a problem for that assumption; viz., since each of (D) and (N) is supposed to include the other in its range, it seems that each must be of higher level than the other, which (in case no one has noticed) is a contradiction. But so far as I can see, this is simply another case like those of  $(\delta)/(\epsilon)$  and the contingent three-of-clubs cycle, in which there is some arbitrariness in the order of evaluation; we get different resolutions depending on whether we treat (D) first or (N) first, but either resolution is satisfactory. I have already argued that we can live with arbitrariness of that sort.

## 6. REVENGE?

But of course: Let  $(L_R)$  be “ $(L_R)$  is not true in any language.”  $(L_R)$  is straightforwardly paradoxical. But our question is, can it be used to derive a contradiction from the truth theory, once our Tarskian stratification is in place? To contradict  $\sim(\exists L)T[(L_R), L]$ , we should have to prove something of the form, “ $T[(L_R), L]$ .”

“ $T[(L_R), O]$ ” is obviously false, since  $(L_R)$  is not grammatical in  $O$ . What about “ $T[(L_R), M_1]$ ”? It is not clear how that formula is to be interpreted. For  $(L_R)$  to be true-in- $M_1$ ,  $(L_R)$  would have to be a sentence of  $M_1$ .  $(L_R)$  does not mention any other language by name, but  $(L_{R\approx 1})$  cannot be about  $M_1$  itself. Is it, then, about other, higher-ordered languages?

I think not. Here is one way to see why. The other languages that concern us are just those in our Tarskian hierarchy, and  $(L_R)$  has a consequence regarding those in particular:  $\sim T[(L_R), O] \& \sim T[(L_R), M_1] \& \sim T[(L_R), M_2] \& \dots$ . In fact,  $(L_R)$  is equivalent to that conjunction itself conjoined with whatever other languages may be in the range of its quantifier. This means, given stratification, that the truth predicate occurring in  $(L_R)$  must be restricted in its application to the highest-ordered of the metalanguages  $M_1, M_2, \dots$ . But of course there is no highest one in Plato’s Heaven. It seems to follow that, contrary to appearances,  $(L_R)$  cannot be applied to any particular lower-ordered premise in order to generate a contradiction. There may *be* a contradiction lurking somewhere, but I see no way of actually deducing one from the Davidsonian truth theory.

In section 4 above I emphasized that “Plato’s Heaven” here is mere abstraction, and that only a few of my metadialects are actually to be found in speakers’ heads. Can we not

then restrict  $(L_R)$  to the latter languages, which quickly come to an end? Let us suppose that there are four of them in the head of a certain speaker at a time. Then  $(L_R)$  is equivalent to  $\sim T[(L_R), O] \& \sim T[(L_R), M_1] \& \sim T[(L_R), M_2] \& \sim T[(L_R), M_3] \& \sim T[(L_R), M_4]$ . So, by our usual assignment policy,  $(L_R)$  must here be  $(L_{R \approx 5})$ . But now, alternately: (i) that is outright contrary to hypothesis; or (ii) we must understand the speaker as quickly growing-in the next higher metadialect, and if (ii), we must then reinterpret  $(L_R)$  as  $(L_{R \approx 6})$ , and we are off and running.

I am not altogether confident of the foregoing position. In particular, there may be a way of deducing a contradiction that I have not seen. Should it be so, we might do well to switch back to Lepore and Ludwig's less conservative method.

There will be other difficulties about semantic predication under my system—more problems created by ambiguity, no doubt, and I have barely mentioned deixis, which is a splitting headache for truth definitions already.<sup>22</sup> But I believe I have made my approach plausible enough to overcome the fear that truth definitions for entire natural languages are doomed to inconsistency.<sup>23</sup>

## References

- Barwise, J., and J. Etchemendy (1987). *The Liar*. Oxford, Oxford University Press.
- Burge, T. (1979). "Semantical Paradox," *Journal of Philosophy* 76, 169-198. Reprinted in Martin (1984), 83-117.
- Davidson, D. (1967). "Truth and Meaning," *Synthese* 17, 304-323. Reprinted in Davidson (2001), 17-36.
- (1969). "On Saying That," in D. Davidson and K.J.J. Hintikka, eds., *Words and Objections*. Dordrecht: D. Reidel. Reprinted in Davidson (2001), 93-108.
- (1973). "In Defense of Convention T," in H. Leblanc, ed., *Truth*, Reprinted in Davidson (2001), 65-76.
- (2001). *Inquiries into Truth and Interpretation*, second edition. Oxford, Oxford University Press.
- Field, H. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.
- Harman, G. (1972). "Logical Form," *Foundations of Language* 9, 38-65.
- Herzberger, H. (1970). "Paradoxes of Grounding in Semantics," *Journal of Philosophy* 67, 145-67.
- Kripke, S. (1975). "Outline of a Theory of Truth," *Journal of Philosophy* 72, 690-716. Reprinted in Martin (1984), 53-81.
- Lepore, E., and K. Ludwig (2005). *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford: Clarendon Press.
- Lycan, W.G. (1984). *Logical Form in Natural Language*. Cambridge, MA, Bradford Books / MIT Press.

- Martin, R., ed. (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford, Oxford University Press.
- McGee, V. (1991). *Truth, Vagueness, and Paradox*. Indianapolis, Hackett Publishing.
- Simmons, K. (1993). *Universality and The Liar*. Cambridge, Cambridge University Press.
- Tarski, A. (1933/1956). "The Concept of Truth in Formalized Languages." In *Logic, Semantics, Metamathematics* (ed. and tr. J.H. Woodger), Oxford, Clarendon Press, 152-197. (Originally published, in Polish, in *Prace Towarzystwa Naukowego Warszawskiego, Wydział III, No. 34* [1933], vii-116.)
- Ulm, M. (1978). "Harman's Account of Semantic Paradoxes," *Studies in Language* 2, 379-383.

### Notes

<sup>1</sup> The term is Tarski's:

A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about a thing at all, we can also speak about it in colloquial language.' (Tarski [1933/1956, p. 164])

It should be noted that the Liar afflicts truth-theoretic semantics of any sort, not just Davidson's. In particular, the problem is no artifact of his extensionalism.

<sup>2</sup> "This idea, we speculate, is what Davidson had in mind in the suggestion, in the passage above, that the claim that natural languages are universal is suspect once we see it leads to paradox" (ibid.). Though I like the idea itself and will vigorously pursue it below, I know of no evidence that Davidson had any type theory even tacitly in mind.

<sup>3</sup> Within a few years, Davidson himself seemed to have given up: "The ideal of a theory of truth for a natural language in a natural language is unattainable if we restrict ourselves to Tarski's methods. The question then arises, how to give up as little as possible, and here theories allowed by Convention T seem in important respects optimal.... It is only the truth predicate itself (and the satisfaction predicate) that cannot be in the object language" (1973, p. 82).

Notice, however, that this fragment strategy concedes, or could as well concede, that whole natural languages are universal, while in the concluding sentence of the original passage Davidson seems unmistakably to question universality. That is why I see at least one more point in the passage than do Lepore and Ludwig (but for their own interpretation see fn 2 above).

<sup>4</sup> If Davidson ever expressed any opinion of dialetheism, I have not heard, but I cannot imagine he would have had any sympathy for it.

<sup>5</sup> See, e.g., Barwise and Etchemendy (1987), McGee (1991), Simmons (1993), Field (2008), and all the works cited therein.

<sup>6</sup> It would have been nice simply to ban self-reference itself, in view of the trouble it has caused in the analysis of English generally; it is tempting to conclude that self-referential expressions are not really well-formed or grammatical even in English. But as we have just seen, such a ban would not solve the Liar problem. Besides, it is just not true that self-referential expressions in English must be ungrammatical or ill-formed, for some are plainly true sentences, such as “This sentence has five words”; and Kripke (1975, p. 692) argues that “Gödel put the issue of the legitimacy of self-referential sentences beyond doubt; he showed that they are as incontestably legitimate as arithmetic itself.”

<sup>7</sup> I leave open for now the question of whether quotation devices themselves are to be left in the object language once semantical terms have been expelled.

<sup>8</sup> Kripke (1975, p. 695) observes that this stratification is not brutally artificial, but expresses Herzberger's (1970) fairly intuitive notion of "groundedness":

First, we make various utterances, such as 'snow is white', which do not involve the notion of truth. We then attribute truth values to these, using a predicate 'true<sub>1</sub>'. ('True<sub>1</sub>' means—roughly—"is a true statement not itself involving truth or allied notions.") We can then form a predicate 'true<sub>2</sub>' applying to sentences involving 'true<sub>1</sub>', and so on.

<sup>9</sup> Of course (as is noted by McGee (1991, p. 70)), Tarski himself had a further reason for doubting that his hierarchical method could be applied to natural languages: A natural language is universal, hence has maximal expressive power, while the object language is less rich. See again the quotation in fn 1 above.

<sup>10</sup> Ulm (1978) made a more detailed version of this criticism against Harman's (1972, sec. 6) solution to the Liar Paradox. My own approach was originally inspired by Harman's; it also bears an affinity to that of Burge (1979).

<sup>11</sup> Lepore and Ludwig make a related point, p. 135, fn 111.

<sup>12</sup> I have been assuming in this paragraph that whenever (L) is considered as a sentence of a particular metalanguage  $M_n$ , the singular term "(L)" is also disambiguated as denoting either the invariably ill-formed "(L-n) is true-in- $M_n$ " or "(L-n) is true-in- $M_{n-1}$ "; on my Tarskian view, the surface sentence "(L) is true" lacks truth-value until its containing-language

parameter has been set—just as does any other surface sentence containing an *ungesättigt* relative term.

<sup>13</sup> There is a temptation to think that since this is so, (L') is *true* after all: What (L') says is that (L') is not true, and if as I allege, (L') is ill-formed, then (L') is indeed not true, and the Paradox comes crashing back in full force. But the inference is fallacious. “(L’)” purports to name a sentence of some language and also to predicate truth of that same sentence; but if (L') is ungrammatical on grounds of violating our type restriction, then the first premise of the foregoing argument is simply false, for (L') does not *say* anything.

<sup>14</sup> The hierarchical treatment also works for truth-functional variants such as Löb's Paradox: (B) “If (B) is true, then P,” which leads to a proof of an arbitrarily chosen proposition. (B) is not a sentence of *O*; at best it would be a sentence of  $M_1$ . But in  $M_1$  it means “If (B-1) is true-in-*O*, then P.” Though the latter is vacuously true by falsity of antecedent, it is true-in- $M_1$  rather than -in-*O*, and so (without equivocation) it does not afford the Modus Ponens needed to derive “P.”

<sup>15</sup> Though possibly you and I have just been adventitiously caused to do so. I am undecided on that.

<sup>16</sup> Kripke (1975). I am grateful to Max Cresswell for a lengthy discussion on this topic.

<sup>17</sup> Admittedly it is a slight idealization, for there are a few odd cases to be sorted out. E.g., what about sentences grammatically incorporating foreign constituents, such as “You've been to the new library, *n'est pas?*,” or a sentence begun by one speaker but completed (*à la*

Huey, Louie and Dewey) by a fellow speaker but in a different tongue. Then there are sardonic cases of intending one's utterance for two audiences simultaneously; someone *might* utter /ɛmpədɔkliyz liypt/, aiming its English meaning at one hearer but its German meaning at another. Though such cases are interesting, I shall assume that they are marginal and that none of them poses any great obstacle to my Liar project. For two more substantive problems, however, see section 4 below.

<sup>18</sup> I here ignore the question of priority or derivativeness as between truth-bearers. I have addressed that issue in Lycan (1984, pp. 191-193).

<sup>19</sup> I was getting tired of /ɛmpədɔkliyz liypt/.

<sup>20</sup> Proof:

1. ( $\delta'$ ) is true(-in- $M_1$ ) iff ( $\epsilon'$ ) is false-in- $O$ . [Truth definition for  $M_1$ ]
2. ( $\epsilon'$ ) is true(-in- $M_2$ ) iff ( $\delta'$ ) is true-in- $M_1$ . [Truth definition for  $M_2$ ]
- $\therefore$  3. ( $\epsilon'$ ) is true-in- $M_2$  iff ( $\epsilon'$ ) is false-in- $O$ . [1,2]
4. Not: ( $\epsilon'$ ) is false-in- $O$ . [Since ( $\epsilon'$ ) is not a sentence of  $O$ ]
- $\therefore$  5. Not: ( $\epsilon'$ ) is true-in- $M_2$ . [3,4]
- $\therefore$  6. Not: ( $\delta'$ ) is true-in- $M_1$ . [2,5]
- $\therefore$  7. Not: ( $\epsilon'$ ) is false-in- $O$ . [1,6] QED

<sup>21</sup> It would go:

1.  $(\alpha_2'')$  is true(-in- $M_1$ ) iff  $(\beta''')$  is true-in- $O$ . [Truth definition for  $M_1$ ]
2.  $(\beta''')$  is true(-in- $M_2$ ) iff  $((\alpha_1)$  is false-in- $M_1$  or  $(\alpha_2'')$  is false-in- $M_1$ ). [Truth definition for  $M_2$ ]
- $\therefore$  3.  $(\beta''')$  is true-in- $M_2$  iff  $(\alpha_2'')$  is false-in- $M_1$ . [From 1 and 2 by truth-functional logic, since  $(\alpha_1)$  is not false-in- $M_1$ ]
- $\therefore$  4.  $(\alpha_2'')$  is true-in- $M_1$  iff  $(\alpha_2'')$  is false-in- $M_1$ . [1,3 (sic)]

But (obviously) the concluding inference equivocates as between  $(\beta''')$ 's being true in  $O$  and  $(\beta''')$ 's being true in  $M_2$ .

<sup>22</sup> For my own general treatment of ambiguity and deixis in Davidsonian truth definitions, see Lycan (1984, Chs. 2 and 3).

<sup>23</sup> I thank Mary Lycan for a conversation that inspired this paper, David Israel for generous and helpful comments on a very messy draft, and Keith Simmons for extensive and crucial criticism of a later but still rocky one.