Truth and Disquotation*

Richard G Heck Jr

According to the redundancy theory of truth, famously championed by Ramsey, all uses of the word 'true' are, in principle, eliminable: Since 'snow is white' is true if, and only if, snow is white, and 'grass is green' is true if, and only if, grass is green, and so forth, an attribution of truth to an explicitly mentioned sentence can always be replaced by the use of that same sentence. It has, however, become clear that, even if the attribution of truth to an explicitly mentioned sentence is redundant, not all uses of the word 'true' will be eliminable. In particular, truth is sometimes attributed not to sentences explicitly mentioned but to sentences merely indicated. I might not know what Russell just said about baseball, but, having the utmost faith in his honesty and knowledge, I might still insist that, whatever he said, it was true. Other examples involve generalization. Someone might say that everything Clinton said about Whitewater was true, even if she had no idea what he had said. Since we do not know what Russell or Clinton said, we cannot eliminate these uses of 'true'. Of course, in these cases, one could perhaps find out what was said, and so one might regard these uses as *in principle* eliminable. But there are other examples, in which one generalizes over infinitely many sentences, and so in which even that strategy fails: Someone might say that all of the infinitely many axioms of Peano arithmetic are true. There is no obvious way to eliminate the word 'true' from that claim, no matter how loosely we construe the notion of elimination.

So the redundancy theory will not do. Its spirit, however, survives in various sorts of 'deflationary' views of truth. According to these views, what the failure of the redundancy theory shows is simply that the word 'true' serves an important expressive function: Without it, we would be unable to say certain things we can say with it. For example, we would be unable to say what we can now say by uttering:

(1) All of the axioms of Peano arithmetic are true.

*Published in *Synthese* 142 (2004), pp. 317–52.

Still, the deflationist holds, we can see the basic insight of the redundancy theory at work here: Although we cannot eliminate the word 'true' from (1), to utter (1) is, in effect, simply to assert all of the axioms of Peano arithmetic. Ramsey's overlooking this fact—that the word 'true' allows us to express infinite conjunctions (and the like) in a finitary language—was thus his only mistake. And this expressive function, according to deflationism, is the *only* (legitimate) one the word 'true' serves. It does not, in particular, serve any semantic function of relating word to world: To utter (1) is not to make a semantic claim—say, one about how the sentences that formulate the axioms of PA relate to the world—but simply to express one's acceptance of a certain theory.[1]

If so, then, as least as far as attributions of truth to explicitly mentioned sentences are concerned, the redundancy theory is right: Such an attribution is always straightforwardly eliminable in favor of the sentence to which truth is attributed; no more (or less) is said when one says that 'snow is white' is true than that snow is white.[2] Moreover, this strong equivalence between an attribution of truth to a sentence and an utterance of that very sentence is what allows us to use the word 'true' to ascribe truth to sentences not explicitly mentioned. It is why, the deflationist will say, saying that what Russell said was true is not, ultimately, to make any semantic claim. Rather, if what Russell said was 'Ted Williams was the greatest hitter of all time', then saying that what he said is true is, in some sense, just saying that Ted Williams was the greatest hitter of all time. Similarly, to utter (1) is just to assert the various axioms of PA: To say that a given axiom of PA is true is just to assert that axiom; to say that they are all true is, therefore, just to assert all of them.

Note that both these claims—that the word 'true' *can* serve an expressive function and that it can*not* serve any robust semantic one—simply follow from the alleged redundancy of attributions of truth to explicitly mentioned sentences. They follow, that is to say, from the the-

[1] Similar remarks would apply to the notion of denotation. As has become customary, however, I~shall keep my focus here on the notion of truth.

[2] As Field notes, when one says that 'snow is white' is true, one commits oneself to the existence of something to which one does not commit oneself when one says merely that snow is white, namely, the sentence 'snow is white'. I take that not to be a serious issue, Field's remarks being more than sufficient to dispose of it. See Field (1994, pp. 250-1), which is reprinted in **?**.

Since I am going to be very critical of Field's views, let me just say explicitly that it is only because I have learned so much from studying his papers that I can be so critical of his views.

sis that, as Quine put it, 'true' disquotes: Attributing truth to 'snow is white' is just attributing whiteness to snow. It is thus the claim that 'true' is disquotational that is at the foundation of deflationism. Indeed, from it follow two further claims, which are characteristic of much deflationist thought.

First, we have no need of a substantial theory of truth or meaning, because the sort of question that gives rise to philosophical theorizing about truth and meaning is misbegotten. An example of such a question would be: *In virtue of what* is it the case that 'snow is white' true if and only if snow is white? Or: In virtue of what does 'snow is white' mean that snow is white? If attributing truth to 'snow is white' is just attributing whiteness to snow, however, it is inappropriate even to ask this sort of question; it is inappropriate even to ask for the sort of explanation of semantic facts that theories of content attempt to provide. That 'snow is white' is true if and only if snow is white is a consequence of basic facts about how we use the word 'true': No other explanation of 'semantic facts' is required. And that's a good thing, since there is none other to be had.

Second, semantic facts can no more figure in deep explanations of other facts than can the fact that no bachelors are female. Such trivialities have no explanatory force. The notion of truth may appear to play an important role in, say, logic. But, the thought is, in so far as it does play such a role, it does so only because logic makes frequent use of the expressive resources the notion of truth makes available. For example, among the things logic tells us is that all instances of the law of noncontradiction are true, that is, that all sentences of the form '$\neg(A \wedge \neg A)$' are true. That is precisely the sort of thing we would be unable to say in natural language without the word 'true'.[3]

Deflationism, as I am understanding it here, is thus the view that 'true' is disquotational, and so that T-sentences, such as

(2) 'Snow is white' is true if, and only if, snow is white,

are mere trivialities—from which it is supposed to follow that 'true' is just an expressive device, that attributions of truth to sentences make no semantic claims, that theories of content are unnecessary and impossible, and that semantic facts have no explanatory force. Hartry Field has defended this sort of view (Field, 1994). Others who describe themselves as deflationists, though, take the fundamental bearers of

[3] For development of this idea, see Horwich (1990), especially Ch. 5.

truth, to be not sentences—or utterances, or sentences plus contexts, or what have you—but propositions. On this view, the central claim of deflationism is that the proposition expressed by 'It is true that snow is white' is equivalent, in some strong sense, to that expressed by 'Snow is white'. Paul Horwich holds this sort of view Horwich (1990), as does Scott Soames Soames (1999). For Horwich, however, the relation of *expression* that holds between an utterance and a proposition is also to be given a deflationary construal Horwich (1998), whereas, for Soames, it is not. This difference matters, as Field notes. On Horwich's view, T-sentences such as (2) will nonetheless turn out to be trivialities; on Soames's, they will not.[4] Horwich's view thus counts as deflationist, for my purposes here; Soames's does not. In any event, my focus is on how the notion of truth applies to sentences.

There are a number of other views that also count as deflationist for my purposes. For example, a view that took truth properly to be explained in terms of substitutional quantification is also deflationist, in my sense. We may define 'true' using substitutional quantification, as follows:

(3) S is true iff $\Sigma p(S = \text{`}p\text{'} \wedge p)$

Then (2) becomes

(2′) $\Sigma p(\text{`snow is white'} = p \wedge p)$ iff snow is white,

which is obviously a logical truth, a mere triviality, certainly *not* a substantial semantic claim about the sentence 'snow is white'. Similar re-

[4] The reason, in short, is that on any view that takes propositions to be the fundamental bearers of truth, attributions of truth to sentences can then be explained as follows:

(i) S is true iff $\exists p[(S \text{ expresses } p) \wedge p \text{ is true}]$.

For example

(ii) 'snow is white' is true iff $\exists p[(\text{'snow is white' expresses } p) \wedge p \text{ is true}]$.

Now, given

(iii) 'snow is white' expresses that snow is white

we can, assuming that 'snow is white' expresses only *one* proposition, easily argue that

(iv)'Snow is white' is true iff snow is white.

If we regard (iii) as licensed by a disquotational construal of the notion of expression, then (iv) will itself have been given a disquotational construal. If, on the other hand, (iii) is given a robust construal, then (iv) too will thereby be given a robust construal.

marks apply to prosentential theories.[5]

Now, why might one find deflationism attractive? Well, one reason is that T-sentences do seem to have some sort of cannotal status. Consider, for example,

(4) 'There are infinitely many twin primes' is true iff there are infinitely many twin primes,

which is an ordinary, material biconditional: It is true if, and only if, its two sides have the same truth-value. But no one knows what the truth-values of the two sides are, since no one knows whether there are infinitely many twin primes. Yet we do know that (4) itself is true. And plainly, we could not know that unless the truth-values of the two sides were tied together somehow: What better explanation—indeed, what *other* explanation—than that it is, somehow or other, part of the meaning of the word 'true' that it disquotes? If so, it seems only a short step to the view that the immediate acceptability of T-sentences—their universal assertability, so to speak—is a consequence of basic facts about how we use the word 'true', in much the way that the universal assertability of 'Bachelors are unmarried' is a consequence of basic facts about how we use the word 'bachelor'. The universal assertability of T-sentences might, for example, be a consequence of the fact that '"snow is white" is true' is, as Field puts it, fully cognitively equivalent to 'snow is white' itself.

Call a set of sentences 'adequate for $\mathcal{L}$' if, for each sentence S, of the language $\mathcal{L}$, it contains exactly one sentence of the form 'S is true iff p'. Any such set, as Tarski observed, fixes the extension of the predicate 'true' on sentences of $\mathcal{L}$ (given, of course, the non-semantic facts). Now, call the language we speak 'English'. By Tarski's observation, the extension of 'true', on sentences of English itself, is fixed by any set of sentences adequate for English. But the T-sentences for sentences of English, as stated in English, are adequate for English, so they fix the extension of 'true' on sentences of English. But if the T-sentences are trivial and uninformative, mere consequences of basic facts about how we use the word 'true', then the extension of the word 'true' on sentences of English is fixed by trivialities.[6]

[5] See Grover et al. (1975). It is less clear to me whether these remarks apply to Jody Azzouni's account in **?**. But then, it is also unclear to me whether Azzouni's theory is a deflationist one.

[6] No collection of trivialities, stated in English, fixes the extension of the English

The foregoing constitutes an argument—not one I would endorse—that our *ordinary* notion of truth is deflationist. Even if that is wrong, however, it might seem that we can always introduce a disquotational notion of truth into ordinary language by stipulating that the T-sentences are to hold, or that '"Snow is white" is true' is to be fully cognitively equivalent to 'snow is white', or what have you. But a notion of truth so introduced will obviously validate all the T-sentences, so it is unclear how it would differ from our ordinary notion of truth. It is, in particular, unclear that there is any work for the ordinary notion of truth to do that could not equally well be done by a disquotational truth-predicate. For this reason, Field urges, we should adopt at least a *methodological* deflationism: "[W]e should assume full-fledged deflationism as a working hypothesis. That way, if full-fledged deflationism should turn out to be inadequate, we will at least have a clearer sense than we now have of just where it is that inflationist assumptions. . . are needed" Field (1994, p. 284).

I am going to argue that we do not need to be methodological deflationists. More precisely, I will argue that we have no need for a disquotational truth-predicate, that the word 'true', as we have it in ordinary language, is not a disquotational truth-predicate, and that it is not at all clear that it is even possible to introduce a disquotational truth-predicate into ordinary language. If so, we have no clear sense how it is even *possible* to be a methodological deflationist. My goal here, let me emphasize, is not to convince a committed deflationist to abandon his or her position. My goal, rather, is to argue, contrary to what many seem to think, that Tarski's observation—that any set of T-sentences for a language fixes the extension of the truth-predicate on that language—does not commit us, and should not even incline us, to deflationism.

The remainder of the paper is organized as follows. I begin, in section 1, by examining an argument, due to Field, that, to make such generalizations as (1), we must use a disquotational truth-predicate: If so, even if we had a non-disquotational truth-predicate, we would still need a disquotational one. I disagree. In section 2, I consider an argument due to Volker Halbach that purports to show that a theory of truth based upon the T-sentences "does not contribute anything to our knowledge of (non-semantical) facts" Halbach (1999, p. 20). And in

word 'true' on sentences of languages other than English. Here, though, it is natural to appeal to translation: A sentence S of some other language falls within the extension of the English word 'true' if, and only if, it is properly translated by some true sentence S^* of English. We shall return to this suggestion.

section 3, I argue that the deflationist attitude towards T-sentences is inappropriate once we look beyond such familiar examples as (2) and consider sentences that exhibit context-dependence, as almost all sentences of natural language do. It will follow that our ordinary notion of truth is not disquotational and that it is not at all obvious how to introduce a disquotational notion into ordinary language. In the final section, I shall gesture in a more positive direction, making a suggestion about the source of our knowledge of T-sentences and the genesis of the concept of truth.

1 Do We Need a Deflationary Notion of Truth?

It is commonly held that, whether or not we have a concept of truth that is non-disquotational, we clearly *do* have one that is. A characteristic expression of this idea is contained in Field's paper "The Deflationary Conception of Truth". Suppose we have a certain infinitely axiomatized theory, such as the first-order theory of Euclidean geometry, interpreted as a theory of the structure of physical space. Now, suppose I wish to deny this theory, but do not have any particular axiom in mind that I wish to deny. To do so, I might say, 'Not every axiom of this theory is true'. But, says Field, in saying that, what I mean to do is just to deny the (infinite) conjunction of the axioms: I mean to say "something about the structure of space only, not involving the linguistic practices of English speakers", that is, not anything about how the sentences used to state the axioms relate to the world. Field concludes that "even someone who accepts a notion of correspondence truth needs a notion of disquotational truth. . . in addition" Field (1986, p. 59).

I think this argument is specious. But before I explain why, let me remind us why it is important. Suppose Field is right. Then not only is there such a thing as a 'disquotational notion of truth', ordinary speakers presumably possess such a notion of truth and the word 'true' sometimes expresses it: It does so, for example, when ordinary speakers say things like 'Euclidean geometry is not true'. So, if we think we also possess a *non*-disquotational notion of truth and that the word 'true' sometimes expresses *it*, we are committed to the ambiguity of the word 'true'. That already seems uncomfortable. But worse, if we think this non-disquotational notion has an important explanatory role to play in, say, logic, we shall find ourselves having to defend the claim that, when the word 'true' does occur in logic, it expresses, not the disquotational

notion we all need, but the non-disquotational notion that is in dispute. Field's argument thus threatens to place the burden of proof squarely upon the opponent of deflationism.

Now, I do not deny that we do use sentences containing the word 'true', in ordinary language, to express certain infinitary statements (such as the denial of the infinite conjunction of the axioms of Euclidean geometry), statements we would otherwise find it hard to express in our finitary language. It does not follow, however, that we need a special word to enable us to do just that. It might, in particular, be the case that, although an attribution of truth to a sentence is not fully cognitively equivalent to that sentence (or whatever), and although an attribution of truth to all the axioms of PA is not just an assertion of them all, the notion of truth can still be *used* to assert such an infinite conjunction.

One strategy here would be to hold, as Field himself suggests, that "a denial that the axioms of Euclidean geometry are true in a [non-disquotational] sense could be used to convey the belief that they are not all disquotationally true" (Field, 1994, p. 59).[7] One way to defend this view would be to argue, first, that *modulo* the facts about how English is used, disquotational truth and non-disquotational truth are equivalent and, second, that such facts may be presumed to be common knowledge and so fixed in the context in which 'true' is used in the way we are discussing. Hence, it may be presumed to be common knowledge that disquotational and non-disquotational truth are equivalent in such contexts.

The point can be made more simply, however. To deny that all axioms of Euclidean geometry are true is, in effect, to assert that one of them is *not* true. But to do so is, in effect, to make a claim about space. Consider the Parallel Postulate: To commit oneself to its untruth is, in light of its T-sentence, to commit oneself to denying the Parallel Postulate; it is to commit oneself to the claim that there is a point p, and a line l, such that through p there is not exactly one line parallel to l. Similarly, to deny any other axiom will, in light of its T-sentence, be to commit oneself to some claim about space, namely, that expressed by the axiom's negation. One cannot, that is to say, deny that all of the axioms of Euclidean geometry are true, even in a *non*-disquotational sense, without thereby committing oneself to a claim about the struc-

[7] Field speaks here of 'correspondence truth', but I regard that terminology as tendentious and so have replaced it with my own.

ture of space: Given the T-sentences—which hold both for disquotational and non-disquotational notions of truth—to deny that all of the axioms of Euclidean geometry are true is to commit oneself to the infinite disjunction of the negations of the various axioms.

One might worry, however, that uses of a disquotational notion of truth are somehow buried in the remarks I've just made. It is important, therefore, to realize that Field's argument does not depend upon the fact that Euclidean geometry has infinitely many axioms. Let S be some finite set of sentences, maybe a large one, and suppose I want to assert the conjunction of the sentences in S without actually having to state them all. So I say that all sentences in S are true. I take it that Field would also claim that, in so saying, I may be saying something, not about how the sentences in S relate to the world, or anything about how they are used by English speakers, but something about whatever the sentences in S are about, say, space. To do so, I would need to employ, Field would claim, a disquotational notion of truth. But now we can reason as above and conclude that, if the T-sentences for the sentences in S are presumed to be common knowledge, it can also be presumed to be common knowledge that, in committing oneself to the truth of all of the sentences in S, one commits oneself to their conjunction and so to a claim about space. If so, then by saying that all sentences in S are true, one can communicate their conjunction and so communicate a claim about space.

Indeed, if Field's argument is cogent, it ought to apply to small finite sets, and even to a single sentence explicitly identified. So, for example, it ought to be possible for me to say '"snow is white" is true' without saying anything about "the linguistic practices of English speakers". And Field, of course, holds just that: Otherwise, one could hardly avoid saying something about speakers when attributing truth to all the sentences in some set. But it seems clear that one does not need a disquotational notion of truth for this purpose. If it can be presumed to be common knowledge that 'snow is white' is true iff snow is white, then it can be presumed to be common knowledge that, in committing oneself to the truth of 'snow is white', one thereby commits oneself to the whiteness of snow. If so, then by uttering '"snow is white" is true' one can communicate the proposition that snow is white.

The important thing to note is that the argument here assumes only that (it is common knowledge that) 'snow is white' is *materially equivalent* to '"snow is white" is true', not that it they are equivalent in any stronger sense—say, that they say the very same thing or are

fully cognitively equivalent. These stronger claims—the ones the deflationist wishes to make—play no role in explaining how an utterance of "'snow is white" is true' might communicate the proposition that snow is white. So we have not yet been given reason to suppose that we need a disquotational notion of truth. Field has another argument, however, namely, that in cases like that of denying the truth of Euclidean geometry, "the *belief* that we are trying to convey does not involve [a non-disquotational notion of] truth" Field (1994, p. 59, my emphasis). That is to say, even if we do not need a disquotational notion of truth to convey our disbelief in Euclidean geometry, it seems we *do* need such a notion to *disbelieve* Euclidean geometry.

The point is most easily made with respect to believing an infinitely axiomatized theory. In my view, all axioms of Peano arithmetic are true. In saying so, I mean to be saying something about the natural numbers, not something about the meanings of certain (formal) sentences, let alone about the facts, whatever they may be, in virtue of which those sentences have the meanings they do. Now, I have just argued that, in order to convey this belief about the numbers, I do not need to employ a disquotational notion of truth: I can simply say, as I just did, that all axioms of PA are true, presuming that you know the T-sentences for those sentences, and know that I know them, and so presume that you realize, and know that I realize, that committing myself to the truth of all of the axioms of PA commits me to various claims about the numbers. However, consider my belief that all axioms of PA are true. Can I so *believe* without thereby believing something about how those sentences relate to the world, or about the facts in virtue of which they mean what they do, or what have you? If so—and one would certainly hope so—then, or so Field claims, my *belief* must involve a disquotational notion of truth.

Before I address this argument, let me consider another. Suppose I say: Although not all of the axioms of Euclidean geometry are true, they might have been. In so saying, I mean to be saying something about the structure of space: I mean to be saying that it is a contingent matter what the structure of space is, in particular, that it is not as Euclidean geometry would have it. But, one might worry, there are different ways that a sentence that is not true might have been true. One way is for the facts to have been different; another is for the sentence to have meant something other than what it in fact means. Even 'All bachelors are married' might have been true: It would have been true had 'bachelor' meant *married man* instead of what it now means. But if so, then it appears that having this belief about the contingency of the structure

of space requires a disquotational notion of truth. What I mean to say, and what I believe, is not that the axioms of Euclidean geometry might have been true in virtue of their having meant something other than what they in fact mean, but that they might have been true in virtue of space's having had a different structure.

So that is Field's challenge. What shall we say about it?

As has often been pointed out, the word 'true' occurs in a number of different constructions in ordinary language.[8] We have so far been concentrating on attributions of truth to sentences, but 'true' occurs also, and probably more often, in construction with a complement clause, as, for example, in 'It is true that snow is white'. Suppose now that Bill says something and I say

(5) What Bill said was not true, though it might have been.

In so speaking, I may mean to comment on the contingency of the claim Bill made, and not simply on the fact that the words he used might have meant something else. But if we take what Bill said to be a *sentence*, so that truth is here attributed to a sentence, then one way what Bill said might have been true is for the sentence he uttered to have meant something other than what it actually means. Clearly, though, 'what Bill said' is ambiguous,[9] and what one would ordinarily mean by an utterance of (5) is that *the proposition Bill expressed* might have been true. On that reading, the problem we have been discussing does not arise. If Bill uttered 'Water is NaCl', then even if 'water' meant *salt,* then, although 'Water is NaCl' would have been true, what Bill said still would not have been true, for what Bill said was that water is sodium chloride, and that could not have been true.

[8] Another response might begin by emphasizing familiar points about counterfactual conditionals: When one utters a counterfactual—or, indeed, makes any sort of modal claim—one presumes that certain things remain fixed. This phenomenon is not simply a matter of the 'closest possible world', in some absolute sense. Context may, in particular cases, specify that we are discussing only worlds in which certain things remain as they are: Certain facts may, in this context, be presupposed, for example. And so similarly, if I say that the axioms of Euclidean geometry might have been true, I may be presupposing that their *meanings* remained unchanged. Indeed, to express (or believe) what Field is claiming one needs a disquotational notion to express, one might simply say: The axioms of Euclidean geometry might have been true, even if they still meant what they now mean.

I am inclined to think this response is adequate, at least for some cases, but the one considered in the text is more generally applicable.

[9] We need not worry here about whether it is ambiguous, or polysemous, or what have you.

One might respond by attempting to stipulate that, in the example we are considering, the word 'said' is used with its sentential meaning and so that truth is therefore being attributed to sentences. But, if the word 'said' is so used, then what Bill 'said', in that sense—that is, the sentence he uttered—*would* have been true in a world in which 'water' meant *salt*, or so it seems to me. One can raise the question how, if truth is not disquotational, we can express the contingency of the facts Bill meant to be stating, as opposed to the contingency of his words' semantic properties. But one cannot require that it be expressed using only a sentential notion of truth. Consider again, then, the claim that the axioms of Euclidean geometry might have been true. Frege, infamously, never tired of insisting that axioms are not sentences, but thoughts, or propositions.[10] His tirelessness gets tiresome, and his point often seems terminological. But it serves here to remind us that an 'axiom' can be a sentence, but it can also be what the sentence expresses. If so, then, when one speaks of the 'axioms' of Euclidean geometry, one may be speaking either of certain sentences or of what those sentences express. And it seems to me that, ordinarily, when one makes claims like the one we are discussing—that the axioms of Euclidean geometry might all have been true—what one intends is the propositional reading. One often hears it said, for example, that the axioms of PA are not only true but necessary. But, of course, the sentences that express those axioms might have been false: They might have meant something else. What could not have been false are the propositions those axioms express.

Obviously, we should now reconsider our initial response to Field's argument. The problem, recall, was that, when I say 'Not all axioms of Euclidean geometry are true', I may mean to be saying something about the structure of space, something that has nothing to do with semantics. I argued above that, even if the word 'true', as used here, is being applied to sentences, and even if it is non-disquotational, we can still understand how an utterance of this sentence might be used to communicate a proposition about the structure of space, even if what it literally *says* is something that does not just concern the structure of space. A stronger response is now available, however: When we make such claims, and intend them to concern the structure of space (as we ordinarily do), we will usually be using the word 'axiom' in the propositional sense.

[10] See, for example, **?**.

Let us return, then, to Field's worry about our beliefs. The problem, recall, was that, even if we do not need a disquotational notion of truth to convey our disbelief in Euclidean geometry, we still need such a notion to disbelieve it. Or, to use the other example, I need such a notion if I am to believe that all axioms of PA are true without thereby believing something about the semantics of English. The answer should now be clear: I can do precisely that by believing of the axioms of PA, in the propositional sense, that they are true. What I need to do is believe, not that all of these sentences are true, but that all of these propositions—the ones expressed by the axioms of PA—are true. Note carefully how the content of the belief has been formulated: I *identify* the propositions to which I attribute truth by means of the sentences that express them, but the mode of identification is not part of the content of my belief;[11] the belief in question is *not* that the propositions expressed by the axioms of PA are true (though I may, of course, also believe that). This latter belief also concerns the semantic properties of certain sentences; the former does not.

In summary, then: Reflection on our use of the word 'true'—in particular, on those uses that allow us to express certain infinitary claims—although it might initially seem to do so, ultimately gives us no reason to suppose we have, or need to have, either in natural language or in our conceptual toolkit, a disquotational notion of truth. Appearances to the contrary are caused by inattention to the distinction between what is said and what is communicated and, more importantly, by an exclusive focus on attributions of truth to sentences.

Now, one might respond that all that has been shown is that we must choose between a disquotational notion of truth and a propositional one—and so that the price of avoiding the *ideological* commitment to a disquotational notion of truth is an *ontological* commitment to propositions. I expect that Field would not be dissatisfied with that outcome. (Quine certainly wouldn't.) But this response misconstrues the argument given above. The argument does not assume the existence of propositions and a notion of truth that applies to them but only that the *construction* 'It is true that *p*' is available to us in natural language and that some conceptual analogue is available to us in thought. That this construction exists in natural language is utterly uncontroversial. That some analogue is available in thought is *prima facie* extremely plausible. There is no ontological commitment to propositions here unless use

[11] This contrast is, of course, familiar from Kaplan (1978).

of the construction 'that p' already commits us to the existence of propositions. It is, of course, controversial whether it does so, in any but a pleonastic sense. If it does, then we need propositions anyway; if not, then we do not need them here, either.

But one might conceive the problem a bit differently. If propositions were the fundamental truth-bearers—if the truth of sentences had to be explained in terms of the truth of the propositions they expressed—then a commitment to the existence of propositions would be hard to avoid. And there are general arguments—which we find, for example, in Frege—that propositions must be the fundamental truth-bearers. Most familiar is the idea that truth can only be ascribed to sentences (or utterances) in so far as they *mean* something: So truth must be ascribed primarily to the meaning of the sentence, and only derivatively to the sentence, in so far as it means something that is true.[12] I myself find such arguments hard to evaluate, because hard to understand: The key move in the argument—from the claim that only sentences that mean something can be true, to the claim that truth must be ascribed primarily to what the sentences mean—seems to me a complete *non sequitur*.[13] But there is a related argument that is much easier to understand, that seems to have a good deal of force, and that is particularly troublesome here. This argument is that it is obvious how to explain truth for sentences in terms of truth for propositions, but very unobvious how to go back the other way.[14]

If we have a notion of truth for propositions, we can explain the attribution of truth to sentences in the following familiar way: A sentence is true iff it expresses a true proposition. Formally:

(6) S is true iff $\exists p((S \text{ expresses } p) \wedge p \text{ is true})$.

But how might one explain attributions of truth to *propositions* in terms of attributions of truth to sentences? One could try saying that a proposition is true iff there is a true sentence that expresses it:

(7) p is true iff $\exists S((S \text{ expresses } p) \wedge S \text{ is true})$.

But while that works, to some extent, for actuality, it doesn't work for possibility: It implies that it might have been true that there were no

[12] This argument is the one familiar from Frege: see e.g. **?**, pp. 160-1, op. 30. Related arguments appear in Soames (1999, ch. 1).

[13] See Dummett (1991a) for development of this concern.

[14] See Soames (1999, pp. 18-9) for reflections of roughly this sort.

sentences iff there might have been a true sentence that expressed the proposition that there are no sentences. Nor would it help to read scope differently, so far as I can see. So that is a problem. It is, however, important to be clear just which problem it is. There are really two problems here that tend to get run together. The first is a *metaphysical* problem. It begins with the assumptions (i) that there are both sentences and propositions, (ii) that truth is sensibly predicated of both of them, and so (iii) that this fact should be explained, if possible, in terms of some relation between the truth of sentences and the truth of propositions. The second is a *linguistic* problem. It begins with the observation (i′) that we have, in natural language, constructions both of the form '*S* is true' and of the form 'It is true that *p*'; it accepts, as a methodological principle (ii′) that, unless we have good evidence to the contrary, we should assume that a single word 'true' is being used in both these cases; whence (iii′) we should seek an explanation of the meaning of the word 'true' that unifies these two uses.

Now, if one accepts assumption (i) of the metaphysical problem, then that problem will indeed seem pressing, since claim (ii) is obvious. Moreover, there is an obvious relation between the truth of a sentence and the truth of the proposition that sentence expresses. So truth seems to be not two things but one, and (iii) just states the explanatory burden one who wishes to defend that intuition incurs. But (i) is, to put it mildly, controversial, and, if one rejects it, *there is no* metaphysical problem. And I, as it happens, do indeed reject (i).

The linguistic problem, though, is another matter. I am not going to solve it here, if for no other reason than that solving it requires a semantics for complement clauses, and I recently lost mine. But it is perhaps worth noting how the problem looks from the perspective of a familiar treatment that does not take complement clauses to denote propositions, Davidson's paratactic theory Davidson (1984). According to this theory, the word 'that' heading a complement clause is really a demonstrative denoting the sentence following it. So 'John said that snow is white' means, roughly:

> John said that (↓).
>
> > Snow is white.

Now 'It is true that snow is white' is just a stylistic variant of 'That snow is white is true'. And so it means, roughly:

> That (↓) is true.

Snow is white.

The demonstrative in the first sentence has, as its referent, the second sentence, so, on Davidson's theory, 'That snow is white is true' is true if, and only if, 'Snow is white' is true. The meanings of propositional attributions—sentences of the form 'That p is true'—would therefore have been explained in terms of the meanings of sentential ones, were it not for the many problems Davidson's theory is known to face. (And if there weren't already enough, versions of the objection from modality discussed above will arise here, too.) I conjecture, however, that views developed in the wake of the paratactic account[15] will also yield accounts of propositional attributions that relate them to sentential attributions, in a suitable way, though surely not as neatly as Davidson's theory does.

2 Digression: Truth and Infinite Conjunction

As noted above, it is one of deflationism's characteristic theses that the role of the word 'true', in natural language, is merely expressive. Its presence, on this view, allows us not to relate words to world, but rather to say certain sorts of things that we might not otherwise be able to say. As we saw, there are a number of examples commonly cited in this regard that illustrate how a predicate characterized entirely by the T-sentences—a disquotational truth-predicate—might allow us to express infinite conjunctions and disjunctions in a finitary language.

Volker Halbach offers a refined analysis of this sort of claim Halbach (1999). Much of the interest of the paper lies in Halbach's analysis of what it *means* for an infinite conjunction to be expressed using a sentence that contains a disquotational truth-predicate. Let S be an infinite set of sentences; suppose we want to express the infinite conjunction of the sentences in S. Intuitively, the sentence "Every sentence in S is true" should do so. Now, in what sense might it do so? Halbach considers a model-theoretic explication of the claim, and shows that it would be adequate, but he suggests, reasonably enough, that such an explication in not in the spirit of deflationism: A proof-theoretic account would be better. So let Σ be some theory, in a language $\mathcal{L}$; let $\mathcal{L}^T$ be $\mathcal{L}$ expanded by a one-place predicate T; and let Σ^T be Σ plus the T-sentences

[15] For objections to Davidson's theory, and some gestures in the direction of a repair, see Higginbotham (1986). A more developed alternative is in Larson and Ludlow (1993). For criticism of that view, see Fiengo and May (1996).

for the sentences in $\mathcal{L}$.[16] Then we have the following result, Halbach's Proposition 2:

> Let $\varphi(x)$ be a formula of $\mathcal{L}$ with just 'x' free. Then
>
> $$\Sigma + \{\phi(\ulcorner A \urcorner) \rightarrow A : A \in \mathcal{L}\}$$
>
> and
>
> $$\Sigma^T + \forall x[\phi(x) \rightarrow T(x)]$$
>
> prove the same formulae of $\mathcal{L}$. Indeed, the latter is a conservative extension of the former.

That is to say: The effect of adding all instances of $\phi(\ulcorner A \urcorner) \rightarrow A$[17] to Σ is, as regards formulae of $\mathcal{L}$, the same as adding $\forall x[\phi(x) \rightarrow T(x)]$ plus a disquotational theory of truth. Moreover, a similar result can be proved regarding infinite disjunctions (though the details are messier): These will, in a similar sense, be expressed by sentences of the form: $\exists x[\phi(x) \wedge T(x)]$.

Halbach notes that "[a]n examination of the proof of Proposition 2... shows that the use of the truth predicate can be effectively eliminated in any given proof" of a formula of $\mathcal{L}$, which he claims "allows for a non-realist towards the truth-predicate", one comparable to instrumentalism or formalism. The idea is that Proposition 2 shows that a disquotational truth-predicate allows *only* for the expression of infinite conjunctions and disjunctions: If so, then one might well conclude that "the theory of truth does not contribute anything to our knowledge of (non-semantical) facts", a conclusion that does indeed "leave the disquotationalist in a rather comfortable position" Halbach (1999, pp. 19, 20). But Halbach's position is unstable.

Consider claims like: Nothing John said is true. Such claims have as much right to be regarded as among the things having a truth-predicate allow us to express as claims like: Everything John said is true. Such claims are, in fact, simply the negations of sentences that express infinite disjunctions, as Halbach notes (and exploits in his proof of the analogue of Proposition 2). So, on their own, sentences of this sort pose no real problem to Halbach: Such sentences—including mathematically interesting examples like "Every true Σ_1 sentence of the language of

[16] Of course, we're assuming the availability of a coding mechanism.

[17] The quotation-marks here, and in similar cases, are written with invisible ink, to avoid cluttering the text.

arithmetic is provable in *Q*"—can be regarded as expressing infinite conjunctions or disjunctions, or the negations thereof, in Halbach's sense.

However, the conservativeness results do not extend to the *joint* addition of sentences from these various classes to Σ^T. Here is an example. Consider the following two sentences:

(8) $\forall x[\exists n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner) \rightarrow T(x)]$;

(9) $\forall x(T(x) \rightarrow [\forall n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \rightarrow \neg Bew(n, \text{`}0 = 1\text{'}))]$,

where '$Bew(x, y)$' means, as usual, that *x* is (the Gödel number of) a Σ-proof of the formula (with Gödel number) *y*. (8) says that every sentence of the form '*n* is not a proof of "0=1"' is true; (9), which is equivalent to

(9*) $\forall x \forall n[x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \wedge T(x) \rightarrow \neg Bew(n, \text{`}0 = 1\text{'})]$,

says that, for every *n*, if the sentence saying that *n* is not the Gödel number of a proof of '0=1' is true, then *n* is not a proof of '0=1'. It can be shown that PA^T + (8) + (9) proves Con(PA), but that PA plus the 'instances' of (8) and (9), in the relevant sense, has the same theorems as PA. (See the Appendix for the proof.) It follows that no analogue of Proposition (2) holds for the joint addition of sentences expressing infinite conjunctions and disjunctions.

Even if we restrict attention to the use of the truth-predicate to express infinite conjunctions and disjunctions, and their negations, then, there are claims that can be so expressed that, taken together, do indeed extend our knowledge of non-semantical matters. But there is another, to my mind more serious, worry, namely, that there is no obvious reason why we should or must limit our attention to sentences expressing infinite conjunctions and the like. Consider, for example:

(10) $\forall x \forall y[T(x \bar{\wedge} y) \equiv T(x) \wedge T(y)]$,

where $\bar{\wedge}$ denotes the syntactic operation of conjunction. Does (10) express an infinite conjunction? If so, which one? The only one that seems plausible is the conjunction of all the instances of $A \wedge B \equiv A \wedge B$, but that can't be right. That would equally be expressed by

(11) $\forall x \forall y[T(x \bar{\wedge} y) \equiv T(x \bar{\wedge} y)]$,

or even by

(12) $\forall x \forall y T((x \bar{\wedge} y) \bar{\equiv} (x \bar{\wedge} y))$,

which have very different formal properties. For example, (11) is valid. But (10) and its kin constitute a Tarski-style truth-theory for the language of arithmetic, and such a theory proves the consistency of PA: The content of (10) therefore is not plausibly exhausted by the collection of instances of $A \wedge B \equiv A \wedge B$.

That the presence of the word 'true' in natural language allows us to express certain sorts of claims we could not express without it is utterly uncontroversial: It allows us, for example, to express such claims as (10) and the other clauses of a theory of truth, claims that certainly look as if they are relating word to world. Deflationism therefore desperately needs the thesis that the presence of the word 'true' only provides allows us with certain expressive (or 'logical') resources we would otherwise lack. But, except for Halbach's, I know of no attempt either to give an account of what these expressive resources are nor to argue that, in some well-defined sense, they exhaust the utility of the word 'true'. Halbach is to be commended for his effort—and for the elegance of his arguments—but his account cannot be deemed satisfactory, for it simply omits such truth-theoretic clauses such as (10). Is there any satisfactory way for a deflationist to understand such claims?

3 T-sentences

Deflationism comes in many forms. But in all its forms, it is committed to regarding T-sentences not as making semantic claims about the sentences mentioned on their left-hand sides, but as 'trivial' or somehow 'insubstantial'—as somehow akin to logical or analytic truths, in so far as their assertability is a consequence of facts about the 'logic' of the word 'true', that is, of that fact that 'true' disquotes. More to the point, the deflationist regards the triviality of T-sentences as a consequence of the fact that our notion of truth is characterized by them. That is what makes it such a natural thought that, even if our ordinary notion of truth is not disquotational, such a notion could yet be introduced into ordinary language *via* a stipulation of the T-sentences, which would then characterize it. So let us ask: *Are* T-sentences, as they are understood in ordinary language, trivial in this way? *Could* we introduce a disquotational notion of truth by stipulating the truth of the T-sentences?[18]

[18] I shall waive worries about the liar paradox, and the other semantical paradoxes. I do believe they pose a serious problem for deflationism, but I have never been

Consider again

(2) 'Snow is white' is true if, and only if, snow is white.

As I said before, there is certainly something special about such sentences: No appeal to empirical knowledge seems needed to establish their truth; we seem able to know them purely on the basis of reflection. But it is worth noting, initially, that, in establishing the truth of (2) by reflection, we draw upon information not contained in it: To establish (2) by reflection, one must recognize that the sentence mentioned on the left-hand side is the same as the sentence used on the right—and not just that it is the same sentence, in some orthographic sense, but that it has the very same meaning. Compare, for example,

(13) 'John went to the bank' is true iff John went to the bank.

Is that true? Lacking further information, we are unable to say, and we are certainly unable to say simply on the basis of reflection: It depends upon whether the word 'bank' mentioned on the left is the same word as that used on the right. Nothing in the T-sentence itself tells one whether it is, and the situation is no different with (2). One could have a perfectly good understanding of (2) and yet not realize that the same sentence was both used and mentioned, and so not be in a position to recognize, simply by reflection, that it is true.

One can build such information into the T-sentence in this way:

(14) The sentence on the right-hand side of this very biconditional is true, in the very language I am now speaking, if, and only if, snow is white.

It is at least arguable that the truth of (14) will be completely obvious to anyone who understands it and takes a moment to reflect upon what it says. Similarly, something like[19]

(15) The sentence on the right-hand side of this very biconditional is true, in the very language I am now speaking, and understood as it will be when I utter it, if, and only if, John went to the bank.

sure whether the problem is practical (that is, 'merely technical') or principled. See Glanzberg (2004) for reasons to think it is a problem of principle. Recent work of Field's may also bear upon this matter. [REF]

[19] I am making use here of an idea suggested by Tyler Burge in a different context. See [REF].

also seems obviously correct. So biconditionals like (14) and (15), which we might call *self-referential* T-sentences, arguably can be known purely by reflection.

Self-referential T-sentences raise a surprising puzzle, one that is worth a digression. One famous paradox, the postcard paradox, is illustrated by the following pair of sentences:[20]

(16) (17) is false.

(17) (16) is true.

One can formulate a version of this paradox, in a single sentence, using the machinery used to produce self-referential T-sentences:

(18) The sentence on my right-hand side is true iff the sentence on my left-hand side is false.

What is surprising is that it is *not* a trivial exercise to formalize this version of the paradox nor even to generate formal analogues of (14) and (15), so long as one works in the usual language of artihmetic. In the case of (18), the obvious idea is to diagonalize on

(19) $\exists y \exists z[rhs(x, y) \wedge lhs(x, z) \wedge T(y) \equiv \neg T(z)]$,

where $rhs(x, y)$ and $lhs(x, y)$ are formulae representing 'the right-hand side of *x*' and 'the left-hand side of *x*', respectively. That would, of course, yield a formula *P* that was provably equivalent to:

(20) $\exists y \exists z[rhs(\ulcorner P \urcorner, y) \wedge lhs(\ulcorner P \urcorner, z) \wedge T(y) \equiv \neg T(z)]$.

And one might think that would do the trick. Unfortunately, however, very little reflection upon the proof of the diagonal lemma is required to convince one that *P* itself is no more a biconditional than (20) is, whence it provably has neither a right- nor a left-hand side: For no y do we have $rhs(\ulcorner P \urcorner, y)$, so (20) will be trivially false and, indeed, formally refutable. Since *P*, whatever it may be, is provably equivalent to (20), it too is refutable and so is false rather than inconsistent with the T-scheme, as one would have wanted.

There is a lesson here: Examples like (18) show that the familiar technique of Gödel numbering is not, in fact, adequate for the representation of *truly self-referential* sentences from natural language. It is

[20] The name comes from the following version of the paradox: Imagine a a postcard on one side of which is written "The sentence on the other side of this card is false" and on the other side of which is written "The sentence on the other side of this card is true".

often said that the Gödel sentence for PA says of itself that it is unprovable; that the (formal version of the) liar sentence says that it is false; and so forth. The enlightened realize, of course, that these sentences do not *really* say such things: The Gödel sentence for PA is only *provably equivalent* to a sentence that says of it that it is unprovable; similarly for the liar. But such modes of expression are undeniably convenient, so we slide over the difference. In some cases, however, the difference matters, as (18) demonstrates. We will see another example of this phenomenon below.[21]

To return to the main thread, then, self-referential T-sentences, such as (14) and (15), do have some sort of special status: They can arguably be known entirely on the basis of reflection.[22] But, I now want to argue, it does not follow that the facts they report are trivial or in any way insubstantial. Suppose that Bill believes:

(21) Utterances by Heck of 'snow is white' are true iff snow is white.

That is not at all trivial: No amount of reflection should convince Bill of (21), and he should regard his belief as contingent, just as I would regard the corresponding belief about him as contingent: It is no necessary truth that Bill's utterances of 'snow is white' are true iff snow is white, because, among other reasons, it is not a necessary truth that Bill speaks my language. But, of course, what Bill believes about me—what he would express by uttering (21)—might well be what I believe about myself—what I would express by uttering (14). The fact reported by (14) is therefore also contingent. So (14) is a contingent truth known entirely on the basis of reflection.

Though this situation certainly can seem puzzling, it is by now relatively familiar: What we have here is something not unlike a case of the contingent *a priori*—a case not unlike that of 'I am here now'. That I am here now is, in some sense, something I can know purely by reflection, and yet it is a contingent fact that I am here now: I might have been somewhere else. Similarly, that an utterance by me now of 'snow is white' would be true if, and only if, snow is white is, in some sense, something I can know purely by reflection, and yet it too is a contingent fact: Those words, in my mouth, might have meant something else. Moreover, from the fact that there is a sense in which I can know purely

[21] One easy way to resolve this problem is to add terms for all primitive recursive functions to the language of arithmetic. For a different way, see **?**.

[22] See Heck (b) for more discussion of this puzzling aspect of T-sentences.

by reflection that I am here now it does not follow that I cannot sensibly ask *why* I am here now, expecting an answer other than that, since 'I' refers to the utterer, etc., any utterance of 'I am here now' is true.[23] Similarly, that I can know, purely by reflection, that an utterance by me now of 'snow is white' would be true if, and only if, snow is white need not prevent me from sensibly asking *why* that is so, expecting an answer other than that 'true' disquotes.[24]

The possibility of knowing (14) and its ilk purely on the basis of reflection therefore gives us no reason to suppose that such questions as why 'snow is white' is true iff snow is white cannot sensibly be raised. It gives us no more reason to deny that the constitutive question *in virtue of what* 'snow is white' is true iff snow is white cannot sensibly be raised. Of course, that the why-question is intelligible does not guarantee that the in-virtue-of-question is, and so it does not *follow* that such constitutive (philosophical) questions can sensibly be raised. But I am not arguing that they can: I am only observing that, in so far as self-referential T-sentences are, in some sense, epistemically trivial, it does not follow that they are metaphysically or even conceptually trivial; so it does not follow that the explanatory and constitutive questions can't sensibly be raised. Or, to put the point differently: We have been given no reason to suppose that a non-deflationist cannot accommodate the 'specialness' of T-sentences.[25] How a non-deflationist should accommodate them will be discussed in the next section.

The T-sentences we have been considering so far concern the truth of sentences, that is, sentence *types*. As has often been pointed out, however, sentences are not plausible candidates for the fundamental bearers of truth. Indeed, in natural language, one rarely attributes truth to sentences at all, and for good reason, namely, because most sentences of natural languages are context-dependent: Typically, what is expressed by an utterance of a sentence of natural language varies with the context in which it is uttered; it makes no sense to ask of such a sentence whether it is true. The standard examples are sentences like

[23] Thanks to Vann McGee for this way of putting the point.

[24] Many have been tempted to say that all we can know by reflection is that the sentence "'Snow is white" is true iff snow is white' is true, and in particular that we cannot know, purely by reflection, that 'Snow is white' is true iff snow is white. See Dummett (1991b, pp. 69ff) for one expression of this view. Such a treatment of the contingent *a priori* is familiar from Donnellan (1977). That strengthens the analogy that matters here.

[25] Thanks here to Crispin Wright for pressing this issue.

'You are sleepy'. Only of a particular utterance of this sentence can one sensibly ask whether it is true.[26] The same obviously applies to other indexicals and to demonstratives. And the phenomenon is arguably far more extensive than that. Consider the sentence 'Everyone is on the bus'. Does an utterance of this sentence always, or ever, mean that absolutely *everyone* is on the one and only one bus in the entire world? I think not.[27] An utterance of this sentence will typically mean only that everyone in some contextually-determined group is on the one and only one bus in some contextually-determined place (or whatever). Indeed, context-dependence is nearly ubiquitous. All of the sentences we have been discussing, and almost all of the sentences speakers typically utter, are significantly tensed: Whatever the right story about 'everyone' and 'the bus', the sentence 'Everyone is on the bus' expresses different things on different occasions simply in virtue of the context-dependence of the present tense.

Now suppose Jean says, 'She saw that movie yesterday'. How shall we apply a disquotational truth-predicate to this utterance? What would it mean if I said, 'What Jean said is true'? Except in special circumstances, it would obviously be wrong for *me* to say that Jean's utterance is true iff she saw that movie yesterday. But more generally, there appears to be nothing 'trivial', nothing that can be established 'purely by reflection', that fixes the meaning of my ascription of truth to Jean's utterance, in the way (2) might fix the meaning of an utterance of '"Snow is white" is true'. Some progress could perhaps be made. Maybe one does know, purely on the basis of reflection, such things as that the word 'I' refers to the utterer and that 'yesterday' refers to the day prior to the day of utterance. So, if we consider instead an utterance of 'I ran yesterday', perhaps one can know, purely by reflection, that it is true if, and only if, its utterer ran on the day prior to the day of utterance. But to what does an utterance of a demonstrative, such as Jean's utterance of 'that movie', refer? What is the analogue for demonstratives of these familiar observations about indexicals? I very much doubt there is one. It would not help to say that it refers to the object Jean means to be indicating. Even if that were so, 'the object Jean means to be indicating' is rather too much like 'the object to which Jean is referring' to do a

[26] There is a dispute regarding whether utterances or sentences-plus-contexts or what have you should be taken as basic here. I~lean towards utterances, but the issue should not matter for present purposes.

[27] The contrary view has, of course, been held. For criticism of it, and a defense, Stanley and Szabó (2000) and Bach (2000), respectively.

deflationist much good.

Perhaps surprisingly, the context-sensitivity of my own utterances causes similar problems. I can apply the model illustrated by (15) to my own utterances as follows:

(22) The sentence on the right-hand side of this very biconditional is true, in the very language I am now speaking, and understood as it will be when I utter it, if, and only if, she saw that movie yesterday.

But note that (22) only works, as it were, at the very moment when it is uttered. If I say now that a particular utterance I made yesterday of 'She saw that movie yesterday' was true, appeal to (22) does not help fix what this ascription means. There appears to be nothing trivial, nothing that can be established purely by reflection, that tells me how now to apply the word 'true' to utterances of mine that I am not making *right now*. Self-referential T-sentences like (22) do me surprisingly little good.

These observations are hardly new, and we shall consider in a moment how a deflationist might try to accommodate them. Before we do so, however, I think we should pause to appreciate their force. Deflationism's appeal, it seems to me, is due in large part to Tarski's observation that the extension of the word 'true', on English sentences, is fixed by the T-sentences for English sentences—coupled to the claim that the T-sentences for English sentences, as stated in English, are trivial, uninformative, and not in need of explanation. But even if all that were correct—which I have argued it is not—we can now see that it is of doubtful interest. Applications of the word 'true' to English *sentences* are rare, and the application of the word 'true' to *utterances* of English sentences is not plausibly characterized by trivialities, not even by T-sentences carefully modified to fit the case of utterances.

Consider the following familiar sort of example:

(23) Everything Bill said at last week's meeting was true.

This is the sort of sentence deflationists typically mention when explaining the 'expressive' purpose they take the word 'true' to serve: An utterance of (23) is supposed to be equivalent to the conjunction of the sentences Bill uttered. Now, I have already argued that such an utterance need not employ a disquotational notion of truth but can instead employ a propositional one. But a stronger point is now emerging,

namely, that a disquotational notion of truth will do us almost no good at all here. If the sentences Bill uttered are like most of the sentences speakers utter, they will exhibit a good deal of context-dependence, and 'trivial' T-sentences, even self-referential ones, will not be available for Bill's utterances when (23) itself is uttered. It therefore seems to me that deflationism's appeal depends, to a significant extent, upon the neglect of context-dependence. Even the plausibility of the deflationist's treatment *of his own favored examples*, such as (23), depends upon the neglect of context-dependence—so heavily, it seems to me, that, had we begun by discussing the application of truth to utterances, rather than to sentences, I doubt that the redundancy theory would have seemed sufficiently plausible to be worth saving from its own excesses.

A deflationism that takes propositional ascriptions of truth as basic is no less vulnerable to the arguments just rehearsed—not, that is, if its associated theory of propositions and expression is deflationist as well.[28] The reason is that, while 'M-sentences' such as

(24) 'Snow is white' means that snow is white

and

> (24′) 'Snow is white' expresses the proposition that snow is white

may appear to be trivialities, in much the same way T-sentences are, M-sentences for utterances are non-trivial in much the same way T-sentences for utterances are.

How, then, might a deflationist try to handle ascriptions of truth to utterances? The obvious thought is to take ascriptions of truth to utterances by me now to be basic—these can be explained in terms of self-referential T-sentences such as (22)—and then to fix the meanings of ascriptions of truth to utterances not made by me now by means of a relation between those utterances and ones potentially made by me now. So Jean's utterance of 'She saw that movie yesterday' (or one of my own previous utterances of it) is true iff a related utterance made by me now would be true. What we really need, of course, is just a mapping from utterances not made by me now to *sentences*: Jean's utterance of 'She saw that movie yesterday' is true iff an utterance by me now of a related sentence S would be true. The problem is then to characterize

[28] A view, such as Soames's, that couples a deflationary theory of truth for propositions to a non-deflationist theory of expression is *not* vulnerable to this sort of objection.

this mapping—and to do so without invoking semantic notions, such as truth and reference.

Field's response to the problems context-dependence poses for deflationism amounts to a sophisticated development of this general approach. He begins by defining what it is for a sentence (type) to be true relative to an assignment of values to indexicals, demonstratives, and the like. So, for example, 'She saw that movie yesterday' is true relative to the sequence <Mary, *Brazil*, 21 November 1997> iff 'Mary saw *Brazil* on 21 November 1997' is true, which it is, of course, iff Mary saw *Brazil* on 21 November 1997. He then proceeds to define truth for utterances:

> [T]he obvious thing to say is that an utterance... is disquotationally true (for me, as I understand it) iff the sentence [uttered] is true relative to the values... I regard [it] as appropriate to associate with the indexicals. When I say that I "associate values" with an indexical, of course, what I do is associate a mental occurrence of one of my own expressions... with it. Field (1994, p. 280)

At times, it seems as if the key to Field's position is the phrase "for me, as I understand it", the thought being that one need only define truth under an assignment, not absolutely: An utterance is disquotationally true *for me, as I understand it*, if it is true relative to the assignment of values to indexicals I *regard* as appropriate. However, as Field recognizes, there is a strong intuition that "there will typically be a *correct answer* to the question... who another person was referring to with a particular application of 'she', [an intuition] the deflationist seems unable to accommodate" if she insists that only how I understand the utterance matters Field (1994, p. 280). Moreover, there are natural interpretations of examples such as the familiar (23)—"Everything Bill said at last week's meeting was true"—on which what matters is very much not how I understand what Bill said. If, for example, what is at issue was Bill's honesty, what matters is not how *I* would have understood him then let alone how I would understand him now.

In any event, Field insists that the "internal processing story" to which he alludes in the passage quoted above "does a lot to accommodate" these sorts of concerns. It is with this story in mind that Field switches from talk of associating *values* with an indexical to talk of "associat[ing] a mental occurrence of one of my own expressions": So long as we construe *associating a value with an expression* as *assigning a*

reference to an expression, "talk of the 'correct assignment' [will be] a semantic matter which cannot appear in a deflationary account. . . " Field (1994, p. 279). But if we instead construe associating a value with an expression as a form of *translation*, then the intuition that there is a right value to associate with an utterance of a demonstrative becomes the intuition that there is a *correct translation* of it. And although there is, of course, a question whether the conditions on correct translation can be explicated without appeal to semantic notions, it is, says Field, at least unobvious that they cannot be.[29]

The following remark illustrates how Field thinks the "internal processing story" helps:

> In a typical case where we misinterpret a token of 'she'—where we incorrectly interpret [Jean] as meaning Mary when in fact she meant Sheila—we do so because of false beliefs about [Jean's] internal processing: we think that her token [of 'she'] was connected up to an internal file drawer of thoughts involving terms like 'Mary' when in fact it was connected up to a file drawer involving terms like 'Sheila'. Field (1994, pp. 280-1)

Field is here trying to avoid saying that it is wrong to translate 'she' by 'Mary' because this utterance of 'she' denotes not Mary but Sheila. The idea is that the correct translation of an utterance of 'she' depends upon which other expressions Jean 'connects' with it: But—even waiving the question in what sense ordinary speakers have beliefs about other speakers' internal file drawers—there are serious problems with this suggestion.

It is important to recognize, first, that what one ultimately wants to know in such a case is not how to translate Jean's utterance into Mentalese. Suppose I decide that the best way to translate Jean's utterance

[29] The translation of which Field speaks here is not, of course, between Jean's idiolect and mine but between Jean's idiolect and my language of thought. This proposal fits well with the familiar view that, in general, understanding a natural language is translating between it and one's language of thought. Field does not mention this view, but it is the obvious way for a deflationist to respond to analogues of the problems we've just been discussing that can be raised for the understanding of utterances in general: Field will not want to say that understanding an utterance involves knowing its truth-condition, in any but a deflationary sense; so one might expect him to say that understanding an utterance involves knowing how to translate it into one's language of thought. To some extent, my skepticism about Field's proposal derives from general skepticism about this conception of language comprehension, which confuses use with mention. But I cannot pursue the issue here.

of 'she' into my idiolect is to translate it by 'Sheila'. That does not tell me, by itself, about whom Jean was talking. Of course, Field would say that I know trivially that 'Sheila' refers to Sheila in my dialect of Mentalese, so I can now conclude that Jean was talking about Sheila. Nonetheless, it is important to observe that what I really want to know is about whom Jean was talking or, to put it differently, to whom Jean was referring: How to translate her utterance is of no interest to me, except in so far as it yields information about reference. Translation is thus at best a mere mechanism, one used to generate information about reference. Field, once again, would say that he can acknowledge this point: All we need to get from knowledge of how to translate to knowledge of 'reference'[30] is purely disquotational knowledge. But, as we shall see, this reply subtly neglects the problems context-dependence poses for Field's position.

The difficulty here is that Field is assuming that one can resolve the question how an utterance of a demonstrative should be translated prior to resolving the question to what it referred. Perhaps that is possible in some cases, but it is not terribly plausible that it is *always* possible. I would have thought, on the contrary, that the decision how to translate an utterance of a demonstrative—whether by 'Mary' or by 'Sheila'—typically depends upon a prior decision to whom it referred. Field's "internal processing story" is, obviously, offered in an attempt to show otherwise, but it does not work, for it just starts a regress. Even if Jean does associate her utterance of 'she' with 'Sheila' instead of 'Mary', it does not follow that she was referring to Sheila rather than to Mary unless her term 'Sheila' refers to Sheila rather than to Mary. The question will then arise how I am to translate her term 'Sheila', and similar problems will recur. (Muttering 'holism' does not count as a response.) Moreover, interpreting Jean as referring to Mary need not involve supposing that Jean herself uses the name 'Mary' as a name for Mary. Even if she does, my so interpreting her does not require me to regard her utterance of 'she' as connected with this name: I might think that Jean doesn't know that she is referring to Mary, although she is. But worse, interpreting Jean as referring to Mary when she uttered the word 'she' does not require me to suppose that she associates her utterance of 'she' with *any* other referential device. If Jean says 'That one is valuable', pointing toward a collection of marbles, I can surely misinterpret her as having referred to one of the marbles rather than another without sup-

[30] Note on use of this term.

posing that she associates her utterance of 'that one' with some other expression by means of which she might have referred to it. The contrary supposition again seems to lead to a regress.

This last example poses an even more serious problem for Field, for it is not only that Jean may lack any expression other than a demonstrative by means of which she may refer to the relevant marble, so may I. Lacking any name for that marble, however, the only way for me to translate her utterance will be to translate it by a demonstrative, say, 'that marble'. But it is, of course, not enough for me to decide to use this expression to translate Jean's utterance. The expression itself has no definite meaning until its own contextual variability is resolved. I can resolve it, of course, by 'pointing', mentally or physically, to an object, so the question that remains to be answered is at which object I should point. But that is simply the question to which object Jean was referring. No progress was made, then, when I decided what expression I would use to translate Jean's utterance: I clearly cannot answer the question how I should translate her utterance without first answering the question to which object she was referring.

I fear, however, that the foregoing, however convincing it may seem in its own right, while perhaps able to diminish the appeal of Field's proposal, is liable to leave one thinking that there must somehow be something to it. If one assumes the language of thought hypothesis—and Field is, of course, tacitly appealing to some such thesis)—then it is extremely tempting to suppose that, in general, understanding an utterance involves, or perhaps just consists in, translating it into one's dialect of Mentalese. If there is a language of thought, after all, then to have a belief—to have *any* belief—is to have a sentence of Mentalese that expresses it in one's 'belief box'. That is as true for beliefs about the semantic properties of words as it is for beliefs about snow and grass. So if I believe that Jean was talking about Sheila when she uttered the word 'she', my so believing consists in my having a sentence like

(25) Jean was talking about Sheila when she uttered the word 'she'

in my belief box. And one can certainly see why one might want to say that, in having this sentence in my belief box, I am associating the mental term 'Sheila' with Jean's utterance of 'she'. But that would be a mistake. The belief I hold when I have (25) in my belief box is not about my mental term 'Sheila' but about Sheila herself. To associate the term 'Sheila' with Jean's utterance of 'she' would, rather, be for me to have a sentence like

(26) Jean's utterance of 'she' is to be associated with, or translated by, 'Sheila'

in my belief box.

To suppose that understanding an utterance of a demonstrative just is translating it into Mentalese is thus to suppose that understanding is having a sentence like (26) in one's belief box, that is, believing (or knowing) a proposition like that expressed by (26). That view, however, has no appeal at all: Understanding an utterance is believing (or knowing) something like (25). It, however, doesn't express a relation between words and words: It expresses a relation between words and *things*, a relation of the sort one might reasonably call 'semantical'. Now here again, one might attempt to concede this point and insist that one can infer things like (25) from things like (26) and appropriate disquotational claims. But, as we have seen, not only is there is no simply reason to suppose that beliefs like (25) must or even can always be based upon beliefs like (26), there is excellent reason to deny it. In fact, however, I strongly suspect that neither of the two views just mentioned is what actually makes the view that understanding is translation seem so attractive. Rather, I suspect that its appeal is due to a failure to distinguish between (26) and (25) and so to conflate use of the term 'Sheila' in (25) with mention of it in (26).

I can report, on the basis of personal experience, that it is easy nonetheless to feel compelled to say that having (25) in one's belief box simply must be to associate the terms 'she' and 'Sheila' in some sense or other. I sympathize with anyone who does feel so compelled: The temptation to confuse use and mention here is agonizingly persistent. But the only sense in which (25) associates 'Sheila' with 'she' is a sense in which it also associate 'Jean' with 'she'. If I am going to associate Jean, the woman, with 'she', the word—to associate her with it as someone who has recently uttered it—then I must of course use some expression that denotes her to do so. And so, in that sense, I have thereby associated that expression with the word 'she'. But it is obviously irrelevant that the second term of the relation is a word rather than, say, another person. If I am going to associate Jean with Bill, say, as his friend, then I must of course use some expression that denotes her to do so. But one would hardly suppose that it followed that one can only associate Jean with Bill if one first associates her name with him, and the same is true of associating Jean with 'she'.[31] And there is no more reason one cannot

[31] If one did so suppose, then the question would arise whether I can associate Jean's

associate Shelia with 'she'—as the person about whom Jean was talking when she uttered it—without first associating 'Sheila' with it.

Field's appeal to translation as a substitute for semantics therefore fails. A relation between words and words is simply no substitute for the relation between words and things we naïvely suppose is involved in our understanding of demonstrative utterances.

I conclude, then, that there is no good reason to believe that our ordinary notion of truth is disquotational: Our ordinary uses of the predicate 'is true' cannot be characterized in terms of trivialities. Now, as I said earlier, it is tempting to think that, whether we have a disquotational notion of truth in natural language or not, we could always introduce such a notion by stipulating that all the T-sentences are to hold for it. Certainly I could introduce a disquotational truth-predicate into my own idiolect by stipulating that all self-referential T-sentences are to hold for it, but such a predicate would be of limited utility, applying only to sentences as potentially uttered by me now: The stipulation of self-referential T-sentences would not characterize the application of this predicate to utterances made by others or even to utterances made by me at other times. And any attempt to extend its application would apparently have to appeal either to a relation between words and *things* or to a relation between words and *words*. Relations of the former sort are semantical, and so are not available to the deflationist; relations of the latter sort are versions of a notion of translation, and will not substitute for the relation between words and things we need. I do not, note, claim to have proven that these are the only options, though it is hard to see what other options there might be. I do not even claim to have proven that the latter definitely will not work. I only claim to be justifiably skeptical that it will work and therefore to be justifiably skeptical that it is even possible to introduce a disquotational truth-predicate into natural language.

4 Closing

Part of what makes deflationism attractive is the simple fact that T-sentences are, somehow or other, special. We seem to know things like

name with Bill without first associating a name of her name with him. If I cannot, we have a regress; if I can, then there is no reason to deny I could not simply associate her with Bill directly, without first associating her name.

the familiar[32]

(2) 'Snow is white' is true iff snow is white

purely on the basis of reflection. It can thus fairly be asked, at the very least, what explanation I would offer of the specialness of T-sentences, for, obviously, I cannot say that it is simply part of the meaning of the truth-predicate that it disquotes. Moreover, one might want to know what I might offer by way of a 'substantial' account of the notion of truth. In closing, I shall say a few words about these matters.

I favor, myself, a conception of semantics, according to which our semantic competence rests upon knowledge of what is expressed by such sentences as (2).[33] On this picture, a competent speaker of a natural language, just as s'he tacitly knows (or 'cognizes') theories of h'er language's syntax, morphology, and the like, tacitly knows a semantic theory for h'er language; the operation of computational mechanisms that draw upon the information contained in this theory delivers, when all goes well, conscious knowledge of (the propositions expressed by) T-sentences for sentences of h'er language. As is familiar, such a semantic theory will contain axioms assigning denotation to the primitive expressions of the language and other axioms that characterize how the reference of a complex expression is determined by the references of its parts. Simplifying enormously, then, we have things like:

(27) $\ulcorner\alpha$ is white$\urcorner$ is true iff den(α) is white,

where den(α) is the denotation of the term α. Intuitively, one can think of (27) as encoding the information that the predicate 'white' expresses the property of whiteness. (Whether it is adequate to the task is another question, one we may set aside here, since I mean only to be explaining my view.)

The obvious question to ask, then, is how the truth-predicate, which of course occurs in natural language, should be handled in such a semantic theory. And the most obvious answer, it seems to me, is that the English word 'true' expresses the very concept of *truth* that plays a central role in the semantic theories speakers tacitly know. Formalizing this suggestion, then, we have that where α is a name of a sentence:

(28) $\ulcorner\alpha$ is true$\urcorner$ is true iff den(α) is true.

[32] In what follows, I shall waive the concerns about context-dependence raised in the last section. It should be obvious enough in which ways it could be handled.

[33] For some reasons in favor of this view, see Heck (a).

Indeed, what *other* axiom would one choose to govern 'true'? But if we do take (28) to be the axiom that governs the truth-predicate, that is already adequate to explain how we might know such T-sentences as (2) purely on the basis of reflection.

I claim, that is, that from (28) and certain other obvious such clauses, the truth of (2) is derivable. By the rule for 'iff':

(29) '"snow is white" is true iff snow is white' is true iff ['"snow is white" is true' is true iff 'snow is white' is true].

Now, by the rule for 'true', that is, (28):

(30) '"snow is white" is true' is true iff den('"snow is white"') is true.

Since den('"Snow is white"') = 'snow is white', we have:

(31) '"snow is white" is true' is true iff 'snow is white' is true.

But that is just the right-hand side of (29), so we may conclude:

(32) '"snow is white" is true iff snow is white' is true.

As promised.

As many authors have noted, in so far as we have knowledge of T-sentences purely by reflection, that knowledge seems to extend not just to sentences of whose truth-values we are ignorant but even to sentences of whose *meanings* we are ignorant. To take a familiar example, suppose that 'The borogroves are mimsy' were a meaningful sentence and that I were told by a reliable source that it was one, so that I now knew that it was. Then I could rightly take myself to know, immediately, and purely by reflection, that

(33) 'The borogroves are mimsy' is true iff the borogroves are mimsy

is true. But I would not, in these circumstances, know (33) itself: I would not know *that* 'The borogroves are mimsy' is true iff the borogroves are mimsy. I would not suppose that I had even the slightest idea what (33) actually meant. I might not even possess the relevant concepts. Yet the explanation just offered of our knowledge that (2) is true adapts immediately to this case: The proof that (2) is true requires no appeal to any semantic knowledge about the sentence 'snow is white' itself; besides (28) and some simple logical facts, we appealed only to the clause for the biconditional and some basic facts about quotation marks.

Field, on the other hand, can only find our knowledge of (33)'s truth puzzling: As he emphasizes, on the disquotational account, one can apply the truth-predicate only to sentences one understands Field (2001, p. 250), and we are supposing I have no understanding of 'The borogroves are mimsy' and so, *a fortiori*, no understanding of (33). On my view, on the other hand, what we know is simply that (33) is true, and I see no reason to deny that one can have such knowledge even if one lacks all understanding of the sentence mentioned on the left-hand side. The knowledge derives not from one's understanding of the sentence mentioned but from one's understanding of the word 'true', of the phrase 'if and only if', and of the workings of quotation-marks. And that, it seems to me, is just as it should be.

It is important to understand that the preceding concerns our understanding of the *word* 'true' not our grasp of the *concept* of truth. To know (28), on my view, is to understand the word 'true', in English: But, to know (28), one must also possess the concept of truth. Indeed, on my view, to know (27)—or to know any other axiom of a theory of truth for English—that is, to understand *any* expression of English, or of any other natural language, for that matter—one must possess the concept of truth. If so, then our most fundamental concept of truth is a *semantical* concept, in roughly Tarski's sense, one that applies, in the first instance, to (utterances of) sentences of natural language.

It would be in the spirit of the view, obviously, to regard this semantical notion of truth as innate, but that does not mean there is nothing constructive to be said about its nature. On the contrary, one obvious question to ask is what features a concept would have to have if it were to play the role the concept of truth plays in semantic theory and, indeed, in our understanding and use of natural language. Or, to put the question slightly differently: What can we conclude about the concept of truth simply from the fact that it plays the role it does in our understanding of natural language? Studying such questions is one way of addressing the 'philosophical problem of truth' without defending a 'theory' of truth (in the sense that correspondence and coherence theories are theories of truth).[34] These few remarks, of course, do not constitute a substantive theory of truth: But I do hope they do at least suggest the direction in which one might look for one.

[34] It is not, of course, as if this suggestion is original with me: I take it that Davidson and Dummett, among others, hold some version of this view. For a particularly clear statement, see Wiggins (1980).

One might object, however, that the clause I have proposed as governing the word 'true', namely:

(28) ⌜α is true⌝ is true iff den(α) is true,

is inconsistent. The word 'true' that appears within the corner quotes is, of course, the word of *natural* language whose meaning this axiom purports to characterize. The word 'true' that appears outside the quotes is not: If we imagine (28) written in Mentalese, then the word 'true' appearing outside the quotes is a word of Mentalese, not of English. We may suppose, if we like, that the Mentalese term applies only to (utterances of) sentences of natural language: Since Mentalese is not a natural language, (28) therefore incorporates something like Tarski's distinction between object-language and meta-language. One might therefore have hoped it would be consistent. But, under natural assumptions, it is not.

Let us write the word 'true' occurring outside the quotes 'True', to make it clear that Tarski's distinction is being respected here. Now let 'λ' be an *English* expression naming the *English* sentence 'λ is not true'. What was (28) then becomes:

(28′) ⌜α is true⌝ is True iff den(α) is True,

from which we get:

(34) 'λ is true' is True iff den('λ') is True.

But den('λ') = 'λ is not true', so:

(35) 'λ is true' is True iff 'λ is not true' is True.

If, now, we have the usual clause for negation, that will deliver:

(36) 'λ is not true' is True iff 'λ is true' is not True.

But then, putting the last two lines together:

(37) 'λ is true' is True iff 'λ is true' is not True.

Contradiction. Note that the contradiction depends only upon (28) and the clause for negation, that it does not involve any sloppiness about meta-language and object-language, and that the contradiction arises in the meta-language, not the object-language. It is, however, also essential to the argument that λ actually *name* the sentence 'λ is not true':

If we assume only that λ is provably equivalent to it, which is what the diagonal lemma delivers, the argument does not go through—not without additional assumptions, anyway, such as that if $A \equiv B$ is provable, then A is True iff B is true. This is the promised other example of a case in which the difference between true self-reference and what the diagonal lemma delivers matters.

There are various ways one might avoid this contradiction. One could try reformulating (28′) as a rule of inference. One could keep (28′) and forego the usual clause for negation. Or one might attempt to uncover, within our ordinary ascriptions of truth to utterances, sensitivity to context that would make theclaimed contradiction a harmless equivocation. But I think it is worth pausing, before we start debating the virtues of these strategies, to ask whether we need to avoid this particular contradiction for this particular purpose.[35] The contradiction does not arise in any theory that I am suggesting we theorists ought to accept. It arises, rather, within a theory that I am suggesting we all tacitly know and employ when we use the English word 'true'—a theory it is the semanticist's task to *describe*, but one s'he has no need, *qua* semanticist, to endorse. The fact that the liar paradox is derivable from (28′) and the usual clause for negation does not show that there is anything wrong with the proposal that we tacitly know them. On the contrary: Our tacitly knowing them explains both why the liar paradox arises and why attempts to 'solve' it always seem unsatisfying. The right attitude towards this particular contradiction may well be Tarski's: What it shows is that our use of the word 'true', in English, is fundamentally inconsistent.[36]

[35] Of course, the contradiction engendered by the liar paradox may well need avoiding for other purposes, and it obviously does need to be avoided in, say, model theory.

[36] Thanks to Michael Glanzberg and Panu Raatikainen for comments on earlier drafts of this paper.

I have been thinking about deflationism for several years now—ever since I saw drafts of Paul Horwich's *Truth* as a graduate student—and I have discussed it with more colleagues and students than I can remember. A few people stand out, though: George Boolos, Michael Glanzberg, ystein Linnebo, Charles Parsons, Michael Rescorla, Jason Stanley, and Jamie Tappenden, as well as Paul himself. The discussion in my graduate seminar on truth, in Spring 2000, and in my undergraduate course on truth, in Fall 2001, was most helpful. I thank everyone who participated.

Talks based upon this paper were presented, under the title 'T-Sentences', at several places: In December 2001, at a colloquium in honor of Tarski's centennial, held at Boston University and sponsored by the Center for Philosophy and History of Science; in January 2002, at the University of Western Ontario and the University of Michigan; and in April 2002, at the University of California at Irvine. Thanks to all who

Appendix

Consider the following two sentences:

(8) $\forall x[\exists n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner) \rightarrow T(x)]$;

(9) $\forall x(T(x) \rightarrow [\forall n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \rightarrow \neg Bew(n, \text{`}0 = 1\text{'}))]$.

where $Bew(x, y)$ means that *x* is (the Gödel number of) a PA-proof of the formula (with Gödel number) *y*. We show that PA^T+(8)+(9) proves Con(PA), but that PA plus the 'instances' of (8) and (9) has the same theorems as PA. (Recall that PA^T is PA plus the T-sentences.)

The instances of (8), in the sense of Halbach's Proposition 2, are all sentences of the form:

(8′) $\exists n(\ulcorner A\urcorner = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner) \rightarrow T(\ulcorner A\urcorner)$

It is easy to see that all such instances are already theorems of PA: The antecedent will, in each case, be decidable, and so refutable if false; when *A* is of the appropriate form, the antecedent will be provable in PA, but then so will the consequent, since PA does prove, for *each* natural number, that it is not a proof of '0=1'. Similarly, the instances of (9) are of the form:

(9′) $A \rightarrow [\forall n(\ulcorner A\urcorner = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \rightarrow \neg Bew(n, \text{`}0 = 1\text{'}))]$,

and these too are already provable in PA. In each case, it will be decidable whether A is indeed of the form $\neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})$. If A is not of that form, then $\forall n(\ulcorner A\urcorner \neq \ulcorner Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner)$ will be provable, the consequent will follow logically, and hence so will the relevant case of (9′); and if A is of that form, then we can prove that it is a particular sentence $\neg Bew(\mathbf{k}, \text{`}0 = 1\text{'})$ of that form and so can prove $\forall n(\ulcorner A\urcorner = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \equiv n = \mathbf{k})$, for some *k*, whence the consequent will be provably equivalent to $\neg Bew(\mathbf{k}, \text{`}0 = 1\text{'})$, which is again provable in PA. So, to sum up,

attended for their questions, comments, and encouragement. Especially helpful were Jody Azzouni, John Bell, William Demopoulos, Allan Gibbard, Thomas Hofweber, Kent Johnson, James Joyce, Patricia Marino, Robert May, Jason Stanley, David Velleman, and Peter Woodruff.

Some of the ideas in section 3—particularly those concerning the analogy between T-sentences and putative examples of the contingent *a priori*—appear also in Heck (b), an early version of which I read at the Massachusetts Institute of Technology in January 1997. Thanks to all who were present, especially Alex Byrne, Vann McGee, and Bob Stalnaker, for helpful questions. Comments from Crispin Wright on drafts of that paper have improved portions of this one.

all the 'instances' of (8) and (9) are already provable in PA. Hence PA plus the 'instances' of (8) and (9) has the same theorems as PA.

However, PA^T plus (8) and (9) proves Con(PA) and so is *not* a conservative extension of PA. In fact, no appeal to the T-sentences is even needed here: Even PA + (8) + (9) proves Con(PA). By logic, (8) and (9) imply

$$\forall x[\exists n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner) \rightarrow \forall n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \rightarrow \neg Bew(n, \text{`}0 = 1\text{'}))]$$

which in turn implies[37]

$$\forall x \forall n(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner \rightarrow \neg Bew(n, \text{`}0 = 1\text{'}))$$

and so[38]

$$\forall n[\exists x(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'})\urcorner) \rightarrow \neg Bew(n, \text{`}0 = 1\text{'})]$$

But we can prove

$$\forall n \exists x(x = \ulcorner \neg Bew(\mathbf{n}, \text{`}0 = 1\text{'}))$$

for that just says that there is, for each n, a sentence that says that n is not a proof of '0=1'. Hence:

$$\forall n \neg Bew(n, \text{`}0 = 1\text{'})$$

That is: Con(PA).

References

Bach, K. (2000). 'Quantification, qualification and context: A reply to Stanley and Szabó', *Mind and Language* 15: 262–83.

Davidson, D. (1984). 'On saying that', in *Inquiries Into Truth and Interpretation*. Oxford, Clarendon Press, 93–108.

Donnellan, K. (1977). 'The contingent a priori and rigid designators', *Midwest Studies in Philosophy* 2: 12–27.

Dummett, M. (1991a). 'Frege's myth of the third realm', in *Frege and Other Philosophers*. Oxford, Clarendon Press, 249–262.

[37] The inference here is from $\exists x A(x) \rightarrow \forall x(A(x) \rightarrow B)$ to $\forall x(A(x) \rightarrow B)$.

[38] The inference here is a simple quantifier manipulation, inferring from $\forall x \forall n(A(x, n) \rightarrow B(n))$ to $\forall n(\exists x A(x, n) \rightarrow B(n))$.

—— (1991b). *The Logical Basis of Metaphysics*. Cambridge MA, Harvard University Press.

Field, H. (1986). 'The deflationary conception of truth', in G. MacDonald and C. Wright (eds.), *Fact, Science, and Morality*. Oxford, Blackwell, 55–117.

—— (1994). 'Deflationist views of meaning and content', *Mind* 103: 249–285.

—— (2001). 'Deflationist views of meaning and content', in *Truth and the Absence of Fact*. Oxford, Clarendon Press, 104–140.

Fiengo, R. and May, R. (1996). 'Interpreted logical forms: A critique', *Rivista di Linguistica* 8: 249–374.

Glanzberg, M. (2004). 'Minimalism and paradoxes', *Synthese* forthcoming.

Grover, D., Camp, J., and Belnap, N. (1975). 'The prosentential theory of truth', *Philosophical Studies* 27: 73–125.

Halbach, V. (1999). 'Disquotationalism and infinite conjunction', *Mind* 108: 1–22.

Heck, R. G. (**?**). 'Reason and language'. 22–45.

—— (**?**). 'Use and meaning', in *The Philosophy of Michael Dummett*.

Higginbotham, J. (1986). 'Linguistic theory and davidson's program in semantics', in E. Lepore (ed.), *Truth and Interpretation: Perspetives on the Philosophy of Donald Davidson*. Oxford, Basil Blackwell, 29–48.

Horwich, P. (1990). *Truth*. Oxford, Blackwell.

—— (1998). *Meaning*. Oxford, Clarendon Press.

Kaplan, D. (1978). 'Dthat', in P. Cole (ed.), *Pragmatics*. New York, Academic Publishers, 221–43.

Larson, R. and Ludlow, P. (1993). 'Interpreted logical forms', *Synthese* 95: 305–355.

Soames, S. (1999). *Understanding Truth*. Oxford, Clarendon Press.

Stanley, J. and Szabó, Z. G. (2000). 'On quantifier domain restriction', *Mind and Language* 15: 219–261.

Wiggins, D. (1980). 'What would be a substantial theory of truth?', in Z. van Straaten (ed.), *Philosophical Subjects*. Oxford, Clarendon Press, 189–221.