Semantics for Deflationists

Christopher Gauker

Version of August 11, 2004

1. Introduction

As I understand it, deflationism about truth is essentially two theses, one positive and one negative. The positive thesis is that sentences (or assertions or propositions) ascribing truth are in some sense equivalent to the sentences (or assertions or propositions) to which truth is ascribed. If we take "is true" as a predicate properly attaching to the names of sentences, then the positive thesis is that a sentence of the form **[ϕ] is true** is in some sense equivalent to the sentence ϕ (where **[ϕ]** is the quotation name of the sentence ϕ). The negative thesis is that this equivalence does not have to be explained by analyzing the truth of ϕ in terms of real reference relations between its subsentential components and objects and sets or properties. That is, the truth of a sentence does not have to be explained in terms of reference relations of a kind that we can explicate only in terms of spatio-temporal-causal relations between uses of that expression and the object or kind of object that the expression is said to refer to.

In what sense is **[ϕ] is true** equivalent to ϕ? (I will assume throughout that ϕ belongs to the same language as **is true**.) At the very least, we should be able to say that **[ϕ] is true** and ϕ always have the same truth value. If we assume that every sentence is either true or false, then, in light of the semantic paradoxes, that claim can be decisively refuted. But a deflationist cannot very well allow that sentences may be neither true nor false, because a deflationist cannot very well explain the difference between falsehood and lack of truth. The broader question posed by this question about equivalence is how the deflationist might define logical validity. What is needed is a *semantic* definition, by which I mean a definition that we can use to demonstrate of valid arguments that they are

valid and of invalid arguments that they are invalid (which is not to say that it must provide an algorithm for constructing such demonstrations). That need not in itself spell defeat for deflationism, for we might obtain a semantic definition of logical validity of a kind that is acceptable to a deflationist. So the question becomes, what kind of semantic definition of logical validity would be acceptable to a deflationist?

The usual model-theoretic approach defines a logically valid argument as one that preserves truth in a model, or interpretation. But if we define logical validity in terms of truth on an interpretation in this way, then, I will argue, we will inevitably suppose that there is one special interpretation, the intended interpretation, such that truth on that interpretation is truth simpliciter. The intended interpretation will be that which assigns to each nonlogical constant what it really refers to. Thus the question arises whether conceiving of truth in this way is compatible with deflationism. In particular, we have to ask whether the deflationist can give an adequate account of the pertinent reference relation, one that does not resort to treating reference as a real relation between singular terms and objects and between predicates and properties or sets of *n*-tuples. The answer, I will argue, is negative.

So if we wish to attain the goals of deflationism, we will need some other approach to defining logical validity. Fortunately, there is indeed another way to think about logical validity, one that both explicates the equivalence between [ϕ] **is true** and ϕ and eschews the reduction of truth to reference. It begins with the concept of *a context for a conversation*. The context for a conversation, as here conceived, is something objective, determined by the goals of the conversation in light of the character of the environment in which the conversation takes place. For each grammatical type of sentence, we may define the conditions under which sentences of that type are assertible in a context, deniable in a context, or neither. Logical validity may then be defined as preservation of assertibility in a context. Semantic paradoxes are avoided inasmuch as the inferences that seem to take us from facts about the identity of paradoxical sentences

(such as that λ = "λ is not true" or that α = "α is not assertible in any context") to contradictions prove to be invalid.

2. Deflationism in a bivalent setting

One of the weakest kinds of equivalence that a deflationist might claim for [ϕ] **is true** and ϕ is material equivalence. So the deflationist's claim might be simply that, for certain sentences ϕ, [ϕ] **is true** and ϕ have the same truth value. The thesis of deflationism would be question-begging if formulated as just the thesis that for certain sentences ϕ, [ϕ] **is true** and ϕ have the same *truth value*. But the same conception of equivalence can be expressed without employing the concept of truth. In a bivalent context, [ϕ] **is true** and ϕ will have the same truth value if and only if the corresponding instance of the T-schema, [ϕ] **is true if and only if** ϕ is true. It would again be question-begging for the deflationist to characterize these instances of the T-schema simply as *true*. But instead they can be characterized as in some sense *given*. Thus we arrive at a formulation of deflationism that we can call *T-schema deflationism*: For certain sentences ϕ, the T-sentence, [ϕ] **is true if and only if** ϕ, is *given*.

There are various senses in which an instance of the T-schema might be said to be *given*. It might be given in the sense that, in doing proofs, we are permitted to take that instance as a premise whenever we choose. Or it might be given in the sense that it is *analytic*. In any case, a sentence's being given has to be sufficient for its being true. Moreover, any instance of the T-schema that is in fact true must qualify as given. Again, where our purpose is to define a deflationary theory of truth, it would be question-begging to explicate givenness in terms of truth. Nonetheless, we may take for granted that the instances of the T-schema that are supposed to explicate the T-schema deflationist's conception of the equivalence between [ϕ] **is true** and ϕ will include all and only those instances that really are true.

Taken as the thesis that *all* sentential instances of the T-schema are given, T-schema deflationism is clearly false. Not all instances of the T-schema can be given, because some are plainly not true. That some instances of the T-schema are not true is the conclusion we must draw from the semantic paradoxes. For example, we might have:

λ = "λ is not true".

But

"λ is not true" is true if and only if λ is not true,

is an instance of the T-schema. From the above identity and this T-sentence, we may infer,

λ is true if and only if λ is not true,

which is plainly not true. It is no help to deny the possibility of identities such as the one here claimed between λ and "λ is not true", because we can construct other nontrue instances of the T-schema without relying on such identities. For example, suppose that on side A of a certain notecard we have the sentence "Every sentence on side B is true" and on side B we have, "No sentence on side A is true". Then if all instances of the T-schema are given, we may derive the conclusion that every sentence on side A is true if and only if *not* every sentence on side A is true, which is certainly not true. In showing this we will rely on the fact that the sole sentence on side A is "Every sentence on side B is true", but such identities cannot be forbidden; they will arise by accident despite our best intentions (as Kripke noted in his 1975).

So the T-schema deflationist cannot maintain that *every* instance of the T-schema is given. In a bivalent setting, we will have to allow that some of them are actually false. (Presumably we will want to say, of the nontrue instance concerning λ, that the left-hand side is false, and the right-hand side is true.) That is why I initially formulated the T-

schema deflationist's claim as only the thesis that *certain* instances of the T-schema are given. Formulated in that way, T-schema deflationism is an explicitly vague thesis. So to get a definite thesis, we will need to answer the question, which instances of the T-schema are the ones that are given? I will now argue that there is no good answer to that question so long as we maintain that every sentence is either true or false.

Anyone who claims to have a theory of something has to have some way of communicating to others what the theory says; otherwise we may deem the theory to be *incomprehensible*. So in particular, a T-schema deflationist who claims to have a conception of the equivalence between **[ϕ] is true** and ϕ must have some way of demonstrating to others that he or she really does have it. One way to demonstrate that one is in possession of a theory is to actually state it by means of a finite number of sentences. In just that sense one might have a theory of physics. Another way might be to define a language and a model for that language and then explain that one's theory consists of the sentences of the language that are true in that model. In just that sense we might have a theory comprising all the truths of arithmetic. But the T-schema deflationist cannot claim to possess a theory in either of these ways. The pertinent set of instances of the T-schema is certainly not finite, and it would be question-begging to identify the pertinent instances in terms of truth.

However, there is another way in which a person might make a theory comprehensible, and that is to describe some definite method for generating a list of all of the sentences that comprise the theory. So unless someone can think of some other reasonable criterion for comprehensibility that T-schema deflationism might pass, we may take for granted that the instances of the T-schema that the T-schema deflationist takes to be given must be at least effectively enumerable. If the instances of the T-schema that the T-schema deflationist takes to be given are not at least effectively enumerable, then we may conclude that T-schema deflationism is strictly *incomprehensible*.

In fact, given only a few further assumptions about our language, T-schema deflationism is demonstrably incomprehensible. The additional assumptions are these: I will assume that the sentences of our language can be written using the syntax of the standard languages of formal logic. I will also assume that our language contains a quotation name for every formula in our language, which I will form with square brackets. I also assume that our language contains its own *diagonal predicate*. If $F(v)$ is a formula of our language containing v as its sole free variable and $[F(v)]$ is its quotation-name, then $F([F(v)])$ is the *diagonal* of $F(v)$. Let the diagonal predicate of our language be **D**, so that a sentence of the form **D***ab* means that the sentence of our language that *a* denotes is the diagonal of the formula of our language that *b* denotes. Finally, I assume that our language includes the language of arithmetic.

A well-known observation of Gödel's, sometimes called the diagonal lemma, tells us that, under these conditions, for any formula $F(v)$ of our language containing v as its sole free variable, we can construct a *Gödel-sentence A* for $F(v)$, which is such that A is true if and only if $F([A])$ is true. In particular, the following sentence is such a Gödel-sentence for $F(v)$:

y(**Dy**[**y**(**Dyx** F(**y**))] F(**y**)).

It is evident that this sentence will be true if and only if the following sentence is true:

F([**y**(**Dy**[**y**(**Dyx** F(**y**))] F(**y**))]).

(By virtue of the meaning of **D**, the first of these two sentences is true if and only if **y**(**y** = [**y**(**Dy**[**y**(**Dyx** F(**y**))] F(**y**))] F(**y**)) is true, which is so if and only if the second sentence is true.)

I will now demonstrate (assuming bivalence) that every sentence of the language is materially equivalent to some instance of the T-schema. Let **S** abbreviate some arbitrary sentence of our language, and consider the following formula in particular:

S if and only if y is true.

Substituting this for "*F*(**y**)" in the schematic Gödel-sentence above, we have:

y(**Dy**[**y**(**Dyx S if and only if y is true**)] **S if and only if y is true**).

Let us abbreviate this sentence as **G**. (So in what follows, wherever I write "**G**", one should read the above sentence, even when "**G**" occurs between square brackets.) Since **G** is a Gödel-sentence for **S if and only if y is true**, it is true if and only if the following sentence is true:

S if and only if [**G**] **is true**.

Thus, we may reason as follows: By the argument just given, **G** and **S if and only if** [**G**] **is true** are materially equivalent. So suppose **S** is true. In that case, **G** and **S if and only if** [**G**] **is true** will be materially equivalent (assuming bivalence) only if **G** and [**G**] **is true** are both true or both false. So [**G**] **is true if and only if G** will be true. Suppose next that [**G**] **is true if and only if G** is true. So **G** and [**G**] **is true** are both true or both false. In either case, **G** and **S if and only if** [**G**] **is true** will be materially equivalent only if **S** is true. So **S** is true if and only if [**G**] **is true if and only if G** is true. But [**G**] **is true if and only if G** is an instance of the T-schema. So every sentence **S** is materially equivalent to some instance of the T-schema. (Notice that this argument fails if we do not assume bivalence. If **G** and [**G**] **is true** can be neither true nor false, then **G** and **S if and only if** [**G**] **is true** might have the same truth value—the value *neither*—though **S** is true and [**G**] **is true if and only if G** is neither true nor false.)

Say that [**G**] **is true if and only if G** is **S**'s *corresponding instance of the T-schema*. What we have just concluded is that **S** is true if and only if its corresponding instance of the T-schema is true. Moreover, since **S** occurs in its corresponding instance of the T-schema (embedded in quotation marks), we find that for any instance of the T-

schema that in this sense corresponds to some sentence, we can "read off" from it the sentence **S** to which it corresponds.

Suppose that the deflationist can have what he or she needs, an effective enumeration of the instances of the T-schema expressing his theory of truth for our language. Next, consider an arbitrary true sentence ϕ of our language. Since ϕ is true and its corresponding instance of the T-schema has the same truth value as ϕ, we may conclude that ϕ's corresponding instance of the T-schema must be true too, and so it ought to be included in the deflationist's theory, and so it ought to turn up somewhere in the enumeration. So for every true sentence of our language, its corresponding instance of the T-schema ought to show up somewhere in the enumeration. But for each of the sentences in this enumeration, we can easily decide whether or not it is the corresponding instance of the T-schema for some sentence. If it is, then, assuming that every instance of the T-schema in the deflationist's theory of truth is true, we will know that the sentence to which it corresponds is true too. Thus an effective enumeration of the instances of the T-schema expressing the deflationist's theory of truth yields as a dividend an effective enumeration of all truths expressible in our language whatsoever.

For any given sentence we can decide whether or not it is a sentence in the language of arithmetic (since that is just a matter of the vocabulary employed). So our enumeration of the truths of our language yields an enumeration of the truths of arithmetic. But the set of truths of arithmetic is a complete theory; that is, for every sentence in the language of arithmetic, either it or its negation belongs to the theory. So our enumeration of the truths of arithmetic yields an effective procedure for deciding whether or not a sentence in the language of arithmetic is true: For any such sentence, wait until either it or its negation shows up in the enumeration. But as is well known, the truths of arithmetic are not decidable in this way. We may conclude that the set of instances of the T-schema that the T-schema deflationist would like to put forward as an

explication of the equivalence of **[ϕ] is true** and ϕ is not effectively enumerable and, therefore, that T-schema deflationism is incomprehensible.[1]

Deflationism, as I have defined it, treats "is true" as a predicate attaching to the names of sentences. A different kind of deflationism, call it *propositional deflationism*, might be defined as the thesis that every proposition of the following *propositional T-schematic form* is "given":

The proposition that ϕ is true if and only ϕ.

(It is not clear to me what might be meant by saying that a proposition has a certain form, but I pass over that complaint.) Such a theory must also yield a theory of truth for sentences, because we can always say (if we believe in propositions in the first place) that a sentence is true in a context if and only if the proposition that it expresses in that context is true. Thus we should expect propositional deflationism to yield a comprehensible theory of truth for sentences as well, or at least for sentences that express the same proposition in every context, such as the sentences of arithmetic. So if the propositional deflationist's account of truth for sentences has the consequence that every sentence is either true or false, then the present argument will be an argument against propositional deflationism as well. (In this way the present argument can be taken as a criticism of Horwich's attempt to state a theory of truth, since he assumes bivalence (1990, p. 80).)

---

[1] This argument is a slightly improved version of the argument that I gave in my 2001. The idea of selecting the Gödel sentence for the formula **S if and only if y is true** comes from Vann McGee (1992), who also used it in an argument against deflationism.

3. Deflationism without bivalence

Again, one of the weakest kinds of equivalence that we might find between **[ϕ] is true** and ϕ is material equivalence. If we abandon bivalence and allow that sentences might be neither true nor false, then to say that two sentences are materially equivalent is to say that they have the same truth value, either both true, or both false, or both neither true nor false. This idea has a natural formulation in terms of inference rules. If for all ϕ, **[ϕ] is true** and ϕ have the same truth value, then all instances of the following two rules of inference are logically valid:

Semantic Ascent	*Semantic Descent*
ϕ	**[ϕ] is true**
¾¾¾¾¾¾	¾¾¾¾¾¾
[ϕ] is true	ϕ

So in the setting of a three-valued semantics, the thesis that **[ϕ] is true** and ϕ are materially equivalent can be formulated as the thesis that these two inference rules (as well as their contrapositives) are valid. Call this *inference-rule deflationism*. Inference-rule deflationism does not reduce to T-schema deflationism because in a three-valued setting these two inference rules may be valid though not every instance of the T-schema is true. Let **[ϕ] is true** and ϕ be two sentences neither of which is true. Then, on the usual three-valued accounts of the biconditional (e.g., the strong Kleene scheme), the T-sentence, **[ϕ] is true if and only if** ϕ, will not be true.[2]

---

[2] In my 1999 I attributed inference-rule deflationism to Hartry Field. This was based on his definition of cognitive equivalence in his 1994b, note 1. The different definition of cognitive equivalence given in note 2 of Field 1994a entails a form of deflationism at least as strong as T-schema deflationism (as I acknowledged as well in my 1999). Field's comments in the postscript to the reprinting of Field 1994a in Field 2001 suggest that at that time he did not wished to be pinned down to any very definite conception of cognitive equivalence.

Inference-rule deflationism does not immediately generate paradoxes in the way that unrestricted T-schema deflationism did. By means of Semantic Ascent and Semantic Descent, together with certain other inference rules, we can indeed derive contradictions from identities such as λ = "λ is not true". But in a three-valued setting, we can block such derivations by rejecting some of the other inference rules employed in the derivations instead of rejecting Semantic Ascent and Semantic Descent. For example, we might try to derive a contradiction from a plain fact about the identity of the liar sentence as follows:

1. λ = "λ is not true". (A plain fact.)
2. Suppose λ is true.
3. Given 2, "λ is not true" is true. (From 1 and 2, by the laws of identity.)
4. Given 2, λ is not true. (From 3, by Semantic Descent.)
5. λ is not true. (From 2−4, by a form of Indirect Proof.)
6. Suppose λ is not true.
7. Given 6, "λ is not true" is true. (From 6, by Semantic Ascent.)
8. Given 6, λ is true. (From 1 and 7, by the laws of identity.)
9. λ is true. (From 6−8, by Indirect Proof.)
10. λ is true and λ is not true. (From 5 and 9.)

The soundness of this reasoning can be denied without rejecting the pertinent instances of Semantic Descent and Semantic Ascent, by rejecting the pertinent instances of Indirect Proof. (I am taking Indirect Proof to be the principle that if a set of sentences $S \cup \{$**not-**$\phi\}$ implies ϕ, then S implies ϕ, and that if a set of sentences $S \cup \{\phi\}$ implies **not-**ϕ, then S implies **not-**ϕ). Similarly, the derivation of contradictory instances of the T-schema by means of instances of Semantic Ascent and Semantic Descent can be rejected by rejecting the required instances of the rule of Conditional Proof (which says that if a set of sentences $S \cup \{\phi\}$ implies ψ, then S implies **If** ϕ **then** ψ).

In order to block the derivation of contradictions in this way, it is necessary to reject the unrestricted use of Indirect Proof and Conditional Proof. It is precisely the decision to reject bivalence and allow that sentences may be neither true nor false that allows us to do this. Suppose that there are counterexamples to Indirect Proof. That is, suppose there is a set of sentences *S* and a sentence ϕ such that *S* {**not-**ϕ} implies ϕ but *S* does not imply ϕ. Since *S* does not imply ϕ, there is an interpretation *M* of the language such that every sentence in *S* is true on *M* and ϕ is not true on *M*. But then since *S* {**not-**ϕ} implies ϕ, **not-**ϕ cannot be true on *M* either. But if ϕ is false on *M*, then (assuming that we do not say anything unusual about negation), **not-**ϕ is surely true on *M*. So since **not-**ϕ is not true on *M*, ϕ is not false on *M*. So ϕ is neither true nor false on *M*. So we cannot have the advantages of inference-rule deflationism without admitting truth-value gaps.

But what can an inference-rule deflationist say about truth value gaps? Presumably, a deflationist would want to explicate falsehood and lack of truth value in the same manner in which he or she proposes to explicate truth. So just as the deflationist explicates truth by positing an equivalence between **[**ϕ**] is true** and ϕ, a deflationist would explicate falsehood by positing an equivalence between **[**ϕ**] is false** and a sentence not containing **false**, and would explicate **not true** by positing an equivalence between **[**ϕ**] is not true** and a sentence not containing **not true**. The trouble is that there is only one plausible candidate for both jobs, namely, **not-**ϕ. The inference-rule deflationist will be driven to say both that **[**ϕ**] is false** is equivalent to **not-**ϕ and that **[**ϕ**] is not true** is equivalent to **not-**ϕ. But if we say that *both* **[**ϕ**] is false** and **[**ϕ**] is not true** are equivalent to **not-**ϕ, then **[**ϕ**] is not true and not false** will be equivalent to a contradiction, **not-**ϕ **and not-not-**ϕ, contrary to our assumption that it is sometimes true that a sentence is neither true nor false.

Perhaps I am mistaken in assuming that the deflationist must explicate lack of truth value in the same way he or she explicates truth and falsehood. Instead, the

deflationist might define a class of sentences *NP* (for *nonparadoxical*) and restrict the equivalence theses to the sentences in *NP*. If a sentence ϕ is in *NP*, *then* [ϕ] **is true** is equivalent to ϕ. And if a sentence ϕ is in *NP*, *then* [ϕ] **is false** is equivalent to **not-**ϕ. Both [ϕ] **is not true** and [ϕ] **is not false** may be true, but only if ϕ is not in *NP*.[3] But then how are we to characterize the class of sentences in *NP*? We cannot characterize it simply as consisting of the sentences that are either true or false, since that would beg the question of the nature of truth and falsehood. Since it is the paradoxical sentences (at least) that we want to characterize as neither truth nor false, perhaps we could say that *NP* consists in (or is confined to) *nonparadoxical* sentences. But then how are we to characterize the nonparadoxical sentences apart from the concepts of truth and falsehood? We might try: A sentence ϕ is *nonparadoxical* if and only if the assumption that **[ϕ] is true** is equivalent to ϕ does not allow us to derive contradictions from plain facts about the identities of sentences. But that will not work because by this criterion plainly nonparadoxical sentences may have to be counted as "paradoxical" too. Consider a version of the notecard paradox (described in section 2 above) in which side B contains both "No sentence on side A is true" and "The moon is the moon". In that case, even the assumption that " 'The moon is the moon' is true" is equivalent to "The moon is the moon" will play a role in the derivation of a contradiction from plain facts about the identities of sentences (and the fact that the moon is the moon).

JC Beall (2002) has suggested that a deflationist can admit truth value gaps by distinguishing between strong and weak negation. Both kinds of negation take a truth into a falsehood and a falsehood into a truth, but a strong negation takes a sentence that is neither true nor false into a truth, while a weak negation takes a sentence that is neither true nor false into another sentence that is neither true nor false. Suppose **NOT** expresses

---

[3] Such an approach is suggested by Richard Holton (2000), although ultimately he does not endorse it. Holton takes seriously another idea, without endorsing it, that I ignore altogether, viz., that we cannot truly *say*, of the sentences that are neither true nor false, that they are neither true nor false.

strong negation and **not** expresses weak negation. Then the deflationist might say that [ϕ] **is neither true nor false** is equivalent to **NOT**-(ϕ **or not** ϕ), which is true in the case where ϕ is neither true nor false. But suppose that **s** = [**s is NOT true**] is true, so that **s is NOT true** is a liar, and compare these three sentences: (a) **s is NOT true**, (b) [**s is NOT true**] **is true**. (c) [**s is NOT true**] **is NOT true**. Since (a) is a liar, we should regard it as neither true nor false. Since (b) ascribes truth to a nontrue sentence, it is either false or neither true nor false. In either case, because (c) is the strong negation of a nontrue sentence, it is true. But (a) results from (c) by identity substitution; so if (c) is true, (a) should be true too, contrary to our assumption. So identity substitution seems to fail. Moreover, if **s is NOT true** is neither true nor false, then likewise **s is true** should be neither true nor false, because otherwise **s is NOT true** would be either true or false. And if **s is true** is not true, then **s is NOT NOT true** cannot be true either. So [**s is NOT true**] **is NOT true** is true and **s is NOT NOT true** is not true. So we cannot maintain that in general [ϕ] **is NOT true** is equivalent to **NOT**-ϕ. But an inference-rule deflationist should expect those two forms of sentence to be equivalent.[4]

4. Deflationary model theory

In the previous two sections I have criticized two deflationary attempts to formulate the purported equivalence between [ϕ] **is true** and ϕ. Now I want to set aside the question of how to formulate this particular equivalence in order to ask a broader question: How can a deflationist define the class of logically valid arguments? This is a strictly broader

---

[4] While thus arguing, by appeal to strong negation, that a deflationist might regard liar sentences as neither true nor false, Beall, writing with Armour-Garb, has also been arguing that deflationists might regard liar sentences as both true and false (Beall and Armour-Garb 2003). On this view, the present problem does not arise: [ϕ] **is NOT true** will be equivalent to **NOT**-ϕ. Strangely, though, a deflationist who maintains that a sentence is both true and false must apparently hold as well that it is not either true or false (weak negation). That would seem to be a problem.

question if we assume that the equivalence of **[**ϕ**] is true** and ϕ is a special case of logical validity¾ if we assume, that is, that the claim is that both the argument from **[**ϕ**] is true** to ϕ and the argument from ϕ to **[**ϕ**] is true** are logically valid. Perhaps some deflationists have imagined that it would suffice just to declare that certain rules of inference are valid (or, in the case of rules like Conditional Proof) validity-preserving, and that all and only arguments whose conclusions can be derived from their premises by means of those rules are valid. But for several reasons that cannot be right.

First, if an argument is valid, then that fact is not just the effect of stipulation. There must be some explanation of that fact in terms that we can relate to a larger conception of the nature of linguistic communication. Such an explanation is precisely what an explanation in terms of a semantic definition of logical validity purports to be (even if the practitioners of logic do not always hold firmly in mind a conception of the relation between their definition of logical validity and a broader conception of the nature of language). Second, it is not always intuitively obvious which forms of argument are really valid, and to resolve such issues we need the guidance of a conception of semantics. If we want to say that the venerable rule of Indirect Proof is not always validity-preserving, because sometimes *S* {**not-**ϕ} logically implies ϕ, though *S* does not logically imply ϕ, that verdict should not be just an ad hoc move designed to rescue the deflationist's equivalence thesis; it should be a verdict that we can justify in light of a broader conception of the nature of language. Finally, one of the things we want in logical theory is a method by which we can demonstrate of invalid arguments that they are invalid (as well demonstrate of valid arguments that they are valid). To demonstrate invalidity we need to be able to construct counterexamples of some kind and then use our account of the semantic properties of sentences to demonstrate that the premises are true (or whatever) in the counterexample and the conclusion is not true in it. (This may not be an algorithm, for there may be no algorithm for finding a counterexample even when a

counterexample exists.) We cannot demonstrate invalidity just by *failing to derive* the conclusion from the premises using the arguments that we have declared to be valid.

On the usual model-theoretic account, an argument is said to be valid if and only if the conclusion is true in every model, or interpretation, in which the premises are all true. (I use the terms "model" and "interpretation" interchangeably.) In the simplest sort of case, a model, or interpretation, specifies a domain of interpretation and specifies an assignment of objects from this domain to individual terms (the referents of these terms) and an assignment of sets of *n*-tuples of members of this domain to *n*-ary elementary predicates (the extensions of these predicates). Further, we are provided with an account of the conditions under which each grammatical type of sentence in the language is true in a model. If we want to allow that the truth of a sentence is relative to a world or a time or a context, then we will have to take some steps beyond this simplest sort of case. For instance, if we want to allow that truth is relative to a possible world, our models may contain a domain of worlds and our assignments may assign to each predicate a function from worlds onto referents or extensions. But for simplicity, I will assume in what follows that we are dealing with the simplest sort of case.

So one option for the deflationist might be to take over the standard model-theoretic conception of logical validity wholesale, but then attempt a deflationary "interpretation" (construal) of the terminology it employs. That is the approach I will consider in this section. So the kind of deflationism I am contemplating in this section is one that grants that **is true** and **is false** have *extensions.* But the deflationist expects that somehow it is the fact that [ϕ] **is true** is equivalent to ϕ and the fact that [ϕ] **is false** is equivalent to **not**-[ϕ] that determine the extensions they have. It is not the independent fact that **is true** and **is false** have these extensions that explains the equivalences.

On the surface, the usual model-theoretic conception of logical validity seems quite contrary to the deflationist outlook. If we define logical validity as preservation of truth in a model, then it seems almost inevitable that we will suppose that there is one

special model, call it the *intended interpretation*, such that truth simpliciter is truth in that particular model. The intended interpretation will be that model that assigns to each elementary nonlogical constant that to which it *really refers.* Having posited such an intended interpretation, it seems we will then be committed identifying it in spatio-temporal-causal terms in the way that deflationist denies we can do. In this way, the model-theoretic account of logical validity seems to contradict the negative thesis of deflationism.

Here is why the definition of logical validity in terms of truth on an interpretation commits one to the claim that truth simpliciter is truth on one of those interpretations in particular: Suppose a model-theoretic definition of logical validity (for a given language) is the right way to define validity, so that to say that an argument is valid is to say just that for every model in which the premises are true the conclusion is true. And suppose, for a reductio, that there is no model (of the language) such that truth in that model is truth (in that language). Then to say that an argument is valid is to make no claim about truth. In particular, in saying that an argument is valid we not already saying that if the premises are true in fact, then the conclusion is true in fact. In other words, saying that an argument is valid does not imply that it does not happen that the premises are true in fact while the conclusion is false in fact. So it could happen that an argument was valid in the sense that for every model in which the premises were true the conclusion was true, even though the premises were true in fact and the conclusion was not true in fact. But I take it that a necessary condition on the correctness of any definition of logical validity is that, necessarily, if the definition renders an argument valid, then if the premises are true in fact then the conclusion is true in fact as well. Having assumed that the model-theoretic definition is correct and that there is no intended interpretation, we arrive at a contradiction: Necessarily, if an argument is valid and the premises are true, then so is the conclusion. But it could happen that an argument was valid and the premises were true but the conclusion was false. So if the model-theoretic definition of validity is correct,

then, contrary to our supposition, there must be an intended interpretation, that is, a model such that truth in that model is truth.[5]

One finds in several sources an argument that purports to show that even if the class of models over which we quantify in defining logical validity does not include an intended model, still we can be sure that all of the arguments that are valid according to our model-theoretic definition really are valid. The argument seems to originate with Kreisel (1967, pp. 153-154); it is reiterated by Etchemendy (1990, chapter 11; he demands some modifications) and Cartwright (1994, p. 10). We suppose that we have a deductive calculus that is sound and complete with respect to our model-theoretic semantics. We are not concerned that our model-theoretic definition might invalidate *too many* arguments. So we may assume that all arguments that really are valid are also valid by our definition. Moreover, all of the arguments valid by our definition are provable, since our deductive calculus is complete with respect to our model-theoretic semantics. Finally, all of the arguments that are provable are presumably really valid. So we have come full circle and may conclude that the arguments that really are valid are exactly those that are valid according to our definition. But in drawing this conclusion we have nowhere assumed that our models include an intended interpretation. So it seems we can dispense

---

[5] Against this conclusion it might be argued that those who define logical validity in terms of truth on an interpretation cannot possibly imagine that there is one special interpretation such that truth on that interpretation is truth simpliciter. Any such "intended interpretation" would have to assign the set of all ordinals to the predicate "is an ordinal" and the set of all sets to the predicate "is a set", but, as almost everyone takes for granted, there is no set of all ordinals or set of all sets. On the contrary, I would say that the model-theoretic definition does indeed commit one to supposing that there is an intended interpretation, as I have explained, and the fact that, for the reason just given, there cannot be one is a major, outstanding problem that almost everyone just sweeps under the rug. One approach to this problem attempts to construct a set theory that allows a universal set (see Forster 1992). Another strategy, initiated recently in Rayo and Williamson 2003, defines logical validity by means of a second-order quantification over interpretive relations rather than by means of a first-order quantification over models. Although such approaches might get us around the problem of the domain's inevitably being too small, they do not allow us to deny that there is an intended interpretation.

with that assumption. The trouble with this argument lies in the assumption that all of the provable arguments really are valid. This just takes for granted what we were led to doubt. The deductive calculus is sound; so we can be sure that all of the provable arguments are valid *by our definition*. But to infer that all of the provable arguments really are valid, we need to know that all of the arguments valid by our definition really are valid. If we cannot be sure that there is an intended interpretation among those over which we quantify in the definition of logical validity, we cannot be entirely sure of that.

Having reached the conclusion that adopting the model-theoretic conception of logical validity commits us to there being an intended interpretation, that is, a model such that truth in that model is truth, we can take a further step and conclude that we are committed to somehow *identifying* the intended interpretation. It would be plainly unsatisfactory to say that *there is* an intended interpretation but we cannot in any way give an informative account of *which* model the intended interpretation *is* (which is not to say that we have to explicitly specify its interpretation of every term in the language).

Presumably the domain for the intended interpretation includes at least everything that exists. The term assignment of the intended interpretation may be stipulatively defined as that which assigns to each individual term what it really refers to and assigns to each predicate the extension is really has. Using the word "refers" to encompass the relation between predicates and their extensions as well as the relation between individual terms and their referents, we can say that the intended interpretation assigns to each term what it *refers to*. Thus the requirement that the intended interpretation somehow be identified may be reformulated as the requirement that the reference relation somehow be identified. In this way, the model-theoretic conception of logical validity poses the question, "What is the reference relation?" If we have no satisfactory answer to that, then that is reason to doubt the theory of logical validity—the model-theoretic conception—that puts us in the position of needing an answer to that.

Nonetheless, a deflationist might hope to mimic the methods of model-theoretic semantics while defending a deflationist account of the terminology it employs. In particular, a deflationist might agree that if we define logical validity in terms of truth in a model, then there must be one special model, the intended interpretation, that assigns to each nonlogical constant that to which it really refers. But the deflationist might maintain that this appeal to reference relations is not a problem unless it commits us to constructing a substantive account of the reference relation that attempts to explain in a general way, without restriction to any particular language, what relation must obtain between the use of an expression and some object or set of things in the world in order for the expression to refer to that object or set of things in the world. So a deflationist might deny that we need any such general theory of reference, applicable to other languages as well as our own, and may maintain that we can identify the intended interpretation for our own language by means of a deflationary account of reference. For instance, a deflationist might propose that all we need to understand about reference is the following equivalences:

Reference Equivalence Schemata

[*b*] **refers to** *a* :: *b* **is** *a*.

o **belongs to the extension of** [*F*] :: *o* **is** *F*.

o_1, o_2 **belongs to the extension of** [*R*] :: o_1 *R*'**s** o_2.

The deflationist might claim that such equivalences specify the intended interpretation of our language as precisely as we could wish.

One problem is still how to allow a three-valued semantics. Suppose the deflationist proposes to explain truth value gaps in terms of extensions and antiextensions in the usual way. So the deflationist might say that a sentence of the form *Fa* is neither true nor false if the thing that *a* refers to is a member of neither the extension of *F* nor a

member of the antiextension of *F*. But then the deflationist will have to supplement the deflationary account of membership in an extension with a deflationary account of membership in an antiextension and will have to do so in a way that allows that an object may belong to neither the extension nor the antiextension of a predicate. The problem will be to say something about failure to belong to the extension of a predicate that is different from what the deflationist says about membership in the antiextension of the predicate. The deflationist might propose to treat the predicate **belongs to the extension of** [*F*] as equivalent to **is** *F*, and might propose to treat the predicate **belongs to the antiextension of** [*F*] as equivalent to **is not** *F*. But then there does not seem to be anything left to which the deflationist can treat the predicate **does not belong to the extension of** [*F*] as equivalent.

Another problem is that it is doubtful whether we can really think of the intended interpretation as one model among others if we think of the intended interpretation as specified in this way. We cannot likewise give a deflationary account of models other than the intended interpretation. We have to think of the assignments provided by the nonintended interpretations as literally functions from elementary nonlogical constants to objects and sets of *n*-tuples. Accordingly, if we are to think of the nonintended interpretations as alternatives to the intended interpretation, then we will have to think of the assignment provided by the intended interpretation as genuinely a function from elementary nonlogical constants to objects and sets of *n*-tuples, which it apparently will not be on the deflationist's account. For instance we cannot define a function *f* by writing: For all *x*, for all *y*, $f(x) = y$ if and only there exists an *a* such that $[a] = x$ and $a = y$ and [*a*] refers to *a*. We cannot define a function in that way because such a definition employs a nonsensical quantification binding a variable that occurs both within and outside of square bracket quotation marks.

Further, it is doubtful whether we really can use instances of the above Reference Equivalence Schemata to identify the reference relation. Whether we can do that will

depend on how many vocabulary items have to have their reference specified separately, that is, not by means of some kind of recursion. If in order to specify the intended interpretation we have to specify separately the reference of infinitely many different expressions, then no finite amount of drawing of inferences by means of the Reference Equivalence Schemata will ever do it. It is often said that if a language is to be learnable, then the number of vocabulary items that have to have their reference specified separately has to be finite. But learnability considerations notwithstanding, it is for several reasons not very plausible that the number of such vocabulary items in natural languages really is finite. (Think about numerals. Think about "that"-clauses.) In any case, the learnability argument rests on dubious assumptions about what it is that we learn when we learn a language. (For a detailed development of this objection, see my 1999.)

5. Basic context logic

The rest of this paper will develop an alternative conception of semantics that deserves to be called deflationist. This will yield a definition of logical validity that we can use to demonstrate that arguments are or are not valid, as the case may be. Moreover, it will establish an equivalence between **[**ϕ**] is true** and ϕ without falling prey to semantic paradoxes. It will not appeal to reference relations at all; a fortiori it will not appeal to reference relations of a kind for which we require a substantive theory. However, this approach to semantics will employ a semantic vocabulary that I do not expect to be explicable in the deflationary way. In that respect my theory may fall short of what some deflationists have aspired to.

The first step will be to explain what I mean by a *context*. In the tradition stemming from Kaplan (1989), a context is supposed to be a set of values of parameters such as time, speaker, hearer, etc. In the tradition stemming from Stalnaker (1974), a context is supposed to be a set of shared assumptions, or a set of assumptions that the

speaker supposes are shared. In my terminology, a context is neither of these things. For me, a context can more accurately be described as structure built up from simple sentences that captures what is objectively relevant about the environment in which the conversation that the context pertains to takes place. (For criticism of Stalnaker's definition of contexts in terms of shared assumptions, see my 1998 and my 2003, chapter 5.)

An account of contexts in my sense will have two aspects, a formal aspect and a substantive aspect. The formal aspect explains the formal structure of a context. The formal structure that we have to build into a context will depend on the logical devices present in the language. Initially, we will consider only a very simple language with logical symbols for negation and disjunction only. Every time we add something of a logical nature to the language, such as quantifiers or a truth predicate, we will have to complicate the formal account of contexts as well. The substantive account of contexts will explain what it takes for a structure of the pertinent formal kind to be *the* structure of this formal kind relative to which we wish to evaluate the sentences uttered in a given conversation. At each stage in the development of my theory I will be able to give a precise formal account of the pertinent kind of context. However, I will attempt a substantive account only here at the beginning, for the simplest kind of case.

So to begin, let *PL* be a language containing denumerably many *individual terms*, and, for each *n*, countably many *n*-ary *predicates*, as well as the connectives "¬" and " ". (Syntactically, *individual terms* here are what others might call an *individual constants.* I do not want to call them that because there is no presumption here that they constantly *denote.*) As usual, an atomic sentence consists of an *n*-ary predicate followed by *n* individual terms. Say that a *literal* is any sentence that is either an atomic sentence or the negation of an atomic sentence. For such a language, a context can be defined simply as a formally consistent, but not necessarily maximal set of literals, thus:

A *context* for *PL* is a set Γ of literals of *PL* (possibly empty) such that for no sentence ϕ do both ϕ and ¬ϕ belong to Γ.

Given this definition of context, we can formulate the conditions under which each type of sentence is assertible or deniable in a context, thus:

(A0) If ϕ Γ, then ϕ is assertible in Γ.

(A¬) If ϕ is deniable in Γ, then ¬ϕ is assertible in Γ.

(A) If ϕ is assertible in Γ or ψ is assertible in Γ, then (ϕ ψ) is assertible in Γ.

(ACl) Nothing else is assertible in Γ.

(D0) If ¬ϕ Γ, then ϕ is deniable in Γ.

(D¬) If ϕ is assertible in Γ, then ¬ϕ is deniable in Γ.

(D) If ϕ is deniable in Γ and ψ is deniable in Γ, then (ϕ ψ) is deniable in Γ.

(DCl) Nothing else is deniable in Γ.

In terms of assertibility in a context for *PL*, we can define logical validity for arguments in *PL* thus:

If *S* is a set of sentences of *PL* and ϕ is a sentence of *PL*, then the argument having the sentences in *S* as premises and the sentence ϕ as conclusion is *logically valid* if and only if for every context Γ for *PL*, if every member of *S* is assertible in Γ, then ϕ is assertible in Γ too.

So, for example, the inference from {(**Fa** ∨ **Gb**), ¬ **Gb**} to **Fa** is valid: Suppose (**Fa** **Gb**) is assertible in Γ, so that either **Fa** or **Gb** is assertible in Γ, which means that either **Fa** or **Gb** is a member of Γ. Suppose also that ¬ **Gb** is assertible in Γ, so that **Gb** is deniable in Γ; which means that ¬ **Gb** is a member of Γ. Since it cannot happen that **Gb** and ¬ **Gb** are both members of Γ, **Fa** must be a member of Γ, which means that **Fa** is assertible in Γ.

By contrast, the inference from {(**Fa** **Gb**)} to **Fa** is invalid. Let Γ = {**Gb**}. Then (**Fa** **Gb**) is assertible in Γ, but **Fa** is not.

According to the formal account, a context for *PL* is just a consistent set of literals of the language *PL*. The substantive account of contexts for *PL* will explain what it is for such a context to be *the* context pertinent to a conversation, that is, the context relative to which we should judge sentences of *PL* to be assertible or deniable in the conversation. In explaining this, I will assume that conversations have goals. In paradigm cases, these goals will be practical goals such as hunting buffalo, building a house, or cooking a meal. In addition, there may be goals that are themselves linguistic, such as finding an answer to a question. Conversations do not always have real goals, but it will be hard to find a case in which a conversation is not guided at least by feigned goals. Apart from the guidance of goals, a conversation is liable to reduce to a sequence of verbal routines.

I will suppose, moreover, that we can distinguish between courses of action, in pursuit of a goal, that *accord* with a given set of literals and those that do not. I will not be able to define this accordance, but I can give an example. Suppose that our goal is to obtain clean water for cooking. Consider then the following set of sentences (which I am thinking of as close enough, for purposes of illustration, to a set of literals):

> {Water is in the well. The well is next to Namu's house. Water is in the barrel. *This* pail is not clean. *That* pail is clean.}

An action that accords with this set of sentences relative to the goal will be fetching water from the well next to Namu's house using *that* pail, not *this* one. Another action that accords with this set might amount to cleaning *this* pail and then using it to fetch water from the well next to Namu's house.

Given these assumptions, the context pertinent to a conversation may be defined as follows:

The context for a conversation is the set of literals such that:

(i) all courses of action in accordance with it relative to the goal of the conversation are good ways of achieving the goal, and

(ii) no proper subset of that set has this property.

In short, the context is the smallest consistent set of literals such that every action in accordance with it is a good way of achieving the goal. Continuing the example, if we added "This pail is red", then the resulting set of literals would be too big; the addition of that sentence would do nothing to narrow down the class of good ways of achieving the goal. If we removed "This pail is not clean" and "That pail is clean", then some actions in accordance with the resulting set of sentences would not be good ways of achieving the goal, namely, using the dirty pail to get the water.

I am not in a position to give a similar substantive account of contexts for each of the languages that we will encounter in what follows. In each case, however, the objective in developing such an account would be to explain how the contents of the context comprise exactly what is in some sense most *relevant* given the goals of the conversation and the actual circumstances in which the conversation takes place. Contexts, so understood, are objective in the sense that interlocutors may be unaware of or even mistaken about their contents. That can happen because interlocutors may be unaware of the circumstances in which their conversation takes place or because they not very good at determining what is relevant given their goals and the circumstances. Contexts change because goals are either achieved or abandoned in favor of new goals, or because external circumstances change; they do not change because someone decides to speak *as if* the context had a certain content. (For further discussion of this nature of communication and contexts, see chapter 3 of my 2003 or chapter 1 of my forthcoming.)

Suppose now that we want to add quantifiers to the language. Let *QL* be a language like *PL* except that in addition *QL* contains denumerably many individual

variables and the quantifier " ". Suppose, moreover that one of the binary predicates of *QL* (and *PL*) is the identity sign "=". Let $v\phi$ abbreviate $\neg v\neg\phi$ in the usual way. Say that $\phi c/v$ is the result of substituting c for v wherever v occurs free in ϕ. Say that two individual terms c and d are *identity-linked* in a set S if and only if either $c = d$ S or there is a term e such that c is identity-linked to e in S and e is identity-linked to d in S. We define contexts for *QL* as follows:

A *context* Γ for *QL* is a pair B_Γ, N_Γ such that:

(1) B_Γ, the *base*, is a set of literals of *PL* such that:

(a) for all ϕ, ϕ and $\neg\phi$ are not both in B_Γ, and

(b) if for each i, $1 \leq i \leq n$, c_i and d_i are identity-linked, then not both $\phi c_1/v_1 \ldots c_n/v_n$ and $\neg\phi\, d_1/v_1 \ldots d_n/v_n$ are in B_Γ, and

(2) N_Γ, the *domain*, is a nonempty set of individual terms that includes every individual term that occurs in any member of B_Γ (and possibly other individual terms as well).

To the conditions on assertibility and deniability, we will now add conditions for the identity sentences and the quantified sentences, thus:

(A=) If $\phi c/v$ is assertible in Γ and $c = d$ or $d = c$ is assertible in Γ, then $\phi d/v$ is assertible in Γ.

(A) If for some individual term c, $\phi c/v$ is assertible in Γ, then $v\phi$ is assertible in Γ.

(D=) If $\phi c/v$ is deniable in Γ and $c = d$ or $d = c$ is assertible in Γ, then $\phi d/v$ is deniable in Γ.

(D) If for all individual terms c N_Γ, $\phi c/v$ is deniable in Γ, then $v\phi$ is deniable in Γ.

As always, if a sentence is not assertible or deniable in a context by any of the conditions already laid down, then it is neither assertible nor deniable in the context. Logical validity may be defined as before (with "*QL*" in place of "*PL*").[6]

To illustrate the use of this apparatus, let us see how a sentence on the order of "Every *F* is *G*" might be assertible in a context although "Everything is *G*" is not assertible in that context. Suppose $B_\Gamma = \{\mathbf{Fa}, \mathbf{Ga}, \mathbf{Gb}, \neg\mathbf{Fc}\}$, and $N_\Gamma = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. $\mathbf{xGx}$ ("Everything is *G*") is not assertible in Γ, for while **c** is in N_Γ, **Gc** is not assertible in Γ because it is atomic and does not belong to B_Γ. But $\mathbf{x}(\neg\mathbf{Fx}\ \ \mathbf{Gx})$ ("Every *F* is *G*") is assertible in Γ because for all *c* in N_Γ, either $\neg\mathbf{F}c$ is assertible in Γ or $\mathbf{G}c$ is assertible in Γ. As for **a** and **b**, **Ga** and **Gb** are both assertible in Γ because they both belong to B_Γ. As for **c**, $\neg\mathbf{Fc}$ is assertible in Γ since $\neg\mathbf{Fc}$ is in B_Γ.

The resulting logic is not classical, but that is not a bad thing at all. $(\phi\ \ \neg\phi)$ is not assertible in every context, and that is a reasonable result, because it is not relevant in every conversation. If we are planning a picnic, then "Either it will rain or not" might be relevant, at least as a reminder that rain is an issue, but in most contexts it will be perfectly irrelevant. Furthermore, $v\phi$ does not imply $\phi c/v$ for arbitrary terms *c* of the language. That too is a reasonable result. There will be contexts in which we can relevantly assert "Everyone has arrived" even if Vladimir Putin has not arrived (and is not expected to arrive). So "Everyone has arrived" may be assertible in some context even though "Vladimir Putin has arrived" is not assertible in that context. (For detailed discussions of this failure of universal instantiation, see my 1997 and my 2003, chapter 7.) Classical logic (for arguments containing at most finitely many premises) is recoverable by confining the class of contexts to contexts whose bases are maximal (in the sense that for every atomic sentence either it or its negation belongs).

---

[6] This definition has undesirable consequences in the case of infinite sets of premises (e.g., the omega rule is valid). These can be avoided by means of Leblanc's method of rewrite functions (1976).

Before proceeding, let me explain how and to what extent this approach to semantics circumvents the problem of reference. As I observed in section 4, one of the ways in which we are driven to posit a reference relation is by defining logical validity as preservation of truth on an interpretation; for this forces on us the supposition that there is one special interpretation such that truth (simpliciter) is truth on that interpretation, and then we expect to use the concept of reference to characterize the intended interpretation. The present approach to logical validity avoids that particular motive inasmuch as in judging an argument to be logically valid we do not have to entertain a variety of possible interpretations of the language at all.

There is perhaps some danger that when all is said and done, we will have to resort to real reference relations in order to explicate some of the concepts that this alternative approach relies on. In particular, there might be some temptation to appeal to correspondence relations of some kind in identifying the context pertinent to a conversation. We might be tempted to say that the set of literals pertinent to a conversation is that which *describes* the set of facts that are relevant to the conversation. But since we are no longer appealing to any kind of mapping from expressions into objects or sets of *n*-tuples of objects in the semantics itself, it should not seem just obvious that we will have to do it in that way.

Further, we might still have to appeal to some kind of reference relation when we go to give the semantics for indexical and demonstrative expressions. We might say that what varies from context to context for a demonstrative is which object in the world it *refers* to. Here I will not attempt to develop a logic of demonstratives; so that too will remain an open question. However, the fact that speakers can use indexicals and demonstratives to refer to different objects on different occasions is certainly not a sufficient reason to suppose that we have to explicate the semantic properties of sentences in terms of reference relations that hold between each of the elementary nonlogical constants and appropriate referents. Pointing at an object and uttering "That!"

certainly creates some kind of real relation between the utterance of "That!" and the object; but that fact alone does not show that we can explicate truth by assigning a reference to every noun, verb and adjective.

6. Context logic for "is true" and "is false"

Next we want to define conditions of assertibility and deniability in a context for the sentences of a language containing predicates meaning *true* and *false*. So let *WL* (the "W" standing for the German "wahr") be a language like *QL* except that, in addition: (i) For every sentence ϕ of *WL*, *WL* contains a name of ϕ, which we form by putting square brackets around ϕ, thus: $[\phi]$. We will call these names of sentences *sentence terms.* (ii) *WL* contains denumerably many *sentence variables.* Predicates that may be followed by a sentence term or sentence variable at a given place (as well as by an individual term or individual variable at that place) will be said to be *sentential* at that place. (iii) The predicates of *WL* include the one-place predicates "**T**" and "**F**", to be understood as meaning *true* and *false*, respectively. These predicates will be *sentential* in their one and only place. (iv) *WL* contains a *sentential quantifier* "Σ", and if τ is a sentence variable, and $\psi[\phi]/\tau$ (the result of substituting $[\phi]$ for τ wherever τ is free in ψ) is a sentence of *WL*, then $\Sigma\tau\phi$ is a sentence of *WL* too. Let $\Pi\tau\phi$ abbreviate $\neg\Sigma\tau\neg\phi$.

In addition, we will need a language which is just like *PL* (our quantifier-free language), except that in addition it contains all the sentence terms of *WL*. Call this language *PL*+. Suppose (for the sake of later illustrations) that *PL*, and thus *PL*+, contains the predicate "**U**", which we understand as meaning *utters*, and which is sentential in its second place. *PL*+ does not contain **T** or **F** or Σ, except as parts of sentence terms.

We may now define a context for *WL* as follows:

A *context* Γ for *WL* is a triple $B_\Gamma, N_\Gamma, S_\Gamma$ such that:

(1) B_Γ, the *base*, is a set of literals of *PL+* such that:

(a) for all ϕ, ϕ and $\neg\phi$ are not both in B_Γ, and

(b) if for each i, $1 \leq i \leq n$, c_i and d_i are identity-linked, then not both $\phi c_1/v_1 \ldots c_n/v_n$ and $\neg\phi\, d_1/v_1 \ldots d_n/v_n$ are in B_Γ, and

(c) if $\phi \neq \psi$, then $[\phi]$ and $[\psi]$ are not identity-linked in B_Γ, and

(2) N_Γ, the *domain*, is a nonempty set of individual terms that includes every individual term that occurs in any member of B_Γ, and

(3) S_Γ, the *sentential domain*, is a nonempty set of sentence terms that includes every sentence term that occurs in any member of B_Γ (and possibly other sentence terms as well).

Supposing that contexts are now defined in this way, we may add assertibility and deniability conditions for sentences containing the new vocabulary as follows:

(A**T**) If ϕ is assertible in Γ, then **T**$[\phi]$ is assertible in Γ.

(A**F**) If ϕ is deniable in Γ, then **F**$[\phi]$ is assertible in Γ.

(AΣ) If for some sentence ψ, $\phi[\psi]/\tau$ is assertible in Γ, then $\Sigma\tau\phi$ is assertible in Γ.

(D**T**) If ϕ is deniable in Γ, then **T**$[\phi]$ is deniable in Γ.

(D**F**) If ϕ is assertible in Γ, then **F**$[\phi]$ is deniable in Γ.

(DΣ) If for all sentences ψ such that either ψ is assertible or deniable in Γ or $[\psi]$ S_Γ, $\phi[\psi]/\tau$ is deniable in Γ, then $\Sigma\tau\phi$ is deniable in Γ.

(The usual closure clause still applies, allowing these new conditions on assertibility and deniability.)

For example, let us see how a sentence meaning, "Every sentence John uttered is true" can be assertible even when "Every sentence is true" is not assertible but "John uttered some sentence" is assertible. To construct such a context we need to ensure that for each sentence that is either assertible or deniable or named in the sentential domain, if

that sentence is deniable, then it is deniable that John uttered it. Thus, we will have to employ a recursion, such as the following:

$N_\Gamma = \{\mathbf{a}, \mathbf{b}, \mathbf{j}, \mathbf{d}\}$

$S_{\Gamma_0} = \{[\mathbf{Fa}], [\mathbf{Hc}]\}$

$B_{\Gamma_0} = \{\mathbf{Uj}[\mathbf{Fa}], \neg\, \mathbf{Uj}[\mathbf{Hc}], \mathbf{Fa}, \mathbf{Gb}\}$

$S_{\Gamma_{i+1}} = S_{\Gamma_i} \quad \{[\phi] \mid \phi \text{ is deniable in } B_{\Gamma_i}, N_\Gamma, S_{\Gamma_i}\}$

$B_{\Gamma_{i+1}} = B_{\Gamma_i} \quad \{\neg\, \mathbf{Uj}[\phi] \mid \phi \text{ is deniable in } B_{\Gamma_i}, N_\Gamma, S_{\Gamma_i}\}$

$S_\Gamma = \bigcup S_{\Gamma_j}, j \geq 0.$

$B_\Gamma = \bigcup B_{\Gamma_j}, j \geq 0.$

$\Sigma\mathbf{sUjs}$ ("John uttered some sentence") is assertible in Γ because $\mathbf{Uj}[\mathbf{Fa}]$ is assertible in Γ. $\Pi\mathbf{sTs}$ ("Every sentence is true") is not assertible in Γ, because, although $[\mathbf{Hc}]$ is in S_Γ, $\mathbf{T}[\mathbf{Hc}]$ is not assertible in Γ (because $\mathbf{Hc}$ is not assertible in Γ). Still, $\Pi\mathbf{s}(\neg\, \mathbf{Ujs} \quad \mathbf{Ts})$ ("Every sentence John uttered is true") is assertible in Γ, because for every sentence ϕ, if ϕ is assertible or deniable in Γ or $[\phi]$ is in S_Γ, then either $\neg\, \mathbf{Uj}[\phi]$ is assertible in Γ or $\mathbf{T}[\phi]$ is assertible in Γ. (In the case of $\mathbf{Hc}$ and every sentence deniable in Γ, $\neg\, \mathbf{Uj}[\phi]$ is assertible in Γ, and in the case of $\mathbf{Fa}$, $\mathbf{Gb}$ and every sentence assertible in Γ, $\mathbf{T}[\phi]$ is assertible in Γ.)

Logical validity is defined in the same way as before: If *S* is a set of sentences of *WL* and ϕ is a sentence of *WL*, then the argument having the sentences in *S* as premises and ϕ as conclusion is logically valid if and only if for every context Γ for *WL*, if every member of *S* is assertible in Γ, then ϕ is assertible in Γ too. It is apparent that Semantic Ascent and Semantic Descent are valid without qualification. The validity of Semantic Ascent is obvious from the assertibility conditions for sentences of the form $\mathbf{T}[\phi]$. So consider the case of Semantic Descent: Suppose $\mathbf{T}[\phi]$ is assertible in arbitrary context Γ. There are two ways this might happen. Case 1: ϕ is assertible in Γ. Case 2: For some individual term *c*, $c = [\phi]$ (or $[\phi] = c$) is assertible in Γ and $\mathbf{T}c$ is assertible in Γ. But given

that c **=** [ϕ] is assertible in Γ, and given that for any sentence ψ distinct from ϕ, [ψ] and [ϕ] are not identity-linked in Γ, **T**c can be assertible in Γ only if **T**[ϕ] is assertible on some other grounds, which takes us back to case 1. So in either case, ϕ is assertible in Γ.

Still, there are no paradoxical sentences in *WL*. We can readily allow that there may be contexts Γ such that a sentence like "λ = 'λ is not true'", or **a =** [$\neg$ **Ta**], is a member of B_Γ and so is assertible in Γ. But this does not generate any contradiction. We simply find that $\neg$ **Ta** cannot be either assertible or deniable in such a context. (Likewise, **Ta** cannot be either assertible or deniable in such a context.) This is not because no sentence of the form $\neg$ **Ta** can ever be assertible. For instance, $\neg$ **Ta** would be assertible in a context in which both **a =** [**Gb**] and **Gb** were assertible. We cannot duplicate the paradoxical reasoning of section 3, because we cannot rely on Indirect Proof. We find that {**a =** [$\neg$ **Ta**], **Ta**} implies $\neg$ **Ta**, since **a =** [$\neg$ **Ta**], and **Ta** are never assertible together in a single context (so that {**a =** [$\neg$ **Ta**], **Ta**} implies every sentence of the language). However, {**a =** [$\neg$ **Ta**]} does not imply $\neg$ **Ta**, because $\neg$ **Ta** may fail to be assertible (indeed, cannot be assertible) in a context in which **a =** [$\neg$ **Ta**] is assertible.[7]

Notice that the present semantics finds no semantic difference between falsehood and nontruth. That is, **F**[ϕ] and $\neg$ **T**[ϕ] are assertible in exactly the same contexts (those in which ϕ is deniable) and are deniable in exactly the same contexts (those in which ϕ is assertible). So the sentence $\neg$ (**T**[ϕ] **F**[ϕ]) is a contradiction, assertible in no context. The semantics is nonetheless three-valued, because it is a semantics of assertibility in a

[7] Some theorists of truth who have put forward diagnoses of the semantic paradoxes have wished to claim that in some context-relative sense we may be able to assign truth or falsehood to liar sentences (e.g., Simmons 1993). But as JC Beall has pointed out (2001), this should not be the position of a deflationist. A deflationist should happily concede that in the case of a simple predication, such as **Ta** (as opposed to, say, an occurrence of **T** under the scope of a quantifier), there should be some equivalent sentence not containing the truth predicate, and if there is not, then the sentence is meaningless, or at least truth-value-less.

context, not of truth, and sentences may be neither assertible nor deniable in some contexts.

7. Context logic for "assertible" and "deniable"

If I left it at that, I could be accused of cheating. The challenged posed by the semantic paradoxes is not just to construct a consistent artificial language. It is not even to construct a consistent artificial language containing a predicate having a logic that looks like the logic we expect from the predicate "is true". Rather, at the very least, the challenge is to show that we can formulate our semantical theories of languages in terms of a metalanguage that does not in turn lend itself to semantic paradoxes. (As Simmons, 1993, has shown, other authors are not always very scrupulous in facing up to this aspect of the challenge.) Here we are formulating our semantics as concerned with the property of assertibility in a context. So to meet the challenge we must show that a semantics of the same kind can be constructed for a language of the kind that we use in formulating our semantic theories, and we must be able to assure ourselves by means of that semantics that our language does not lend itself to semantic paradoxes.

Let *AL* be a language like *WL* except that, in addition: (i) If ϕ is a sentence of *AL*, then $[\phi]$ is sentence term of *AL*. (ii) *AL* contains denumerably many terms called *context terms* and, correspondingly, denumerably many *context variables*. (iii) *AL* contains the two-place predicates **Asst**(…, …) and **Den**(…, …), and if γ is a context term and α is either a sentence term or an individual term of *AL*, then both **Asst**(α, γ) and **Den**(α, γ) are sentences of *AL* as well. **Asst**$([\phi], \gamma)$ is understood as meaning "ϕ" *is assertible in* γ, and **Den**$([\phi], \gamma)$ is understood as meaning "ϕ" *is deniable in* γ. (These predicates carry parenthesis for clarity.) (iv) *AL* contains the quantifier "§" (call it the *double ess*), and if ω is a context variable and γ is a context term and $\phi\gamma/\omega$ is a sentence of *AL*, then $\S\omega\phi$ is a sentence of *AL*. Let $\P\omega\phi$ abbreviate $\neg\S\omega\neg\phi$.

In addition, we will need a language which is just like *PL* (our quantifier-free language), except that in addition it contains all the sentence terms of *AL*. Call this language *PL‡*.

In terms of *PL‡* we will now define contexts for *AL* inductively. Initially, we define the set of *basic* contexts for *AL*, and then in terms of those we define the rest of the contexts for *AL*.

A *basic context* Γ for *AL* is a quintuple $\langle B_\Gamma, N_\Gamma, S_\Gamma, C_\Gamma, f_\Gamma \rangle$ such that:

(1) B_Γ, the *base*, is a set of literals of *PL‡* such that:

(a) for all ϕ, ϕ and $\neg\phi$ are not both in B_Γ, and

(b) if for each i, $1 \leq i \leq n$, c_i and d_i are identity-linked, then not both $\phi c_1/v_1 \ldots c_n/v_n$ and $\neg\phi\, d_1/v_1 \ldots d_n/v_n$ are in B_Γ, and

(c) if $\phi \neq \psi$, then $[\phi]$ and $[\psi]$ are not identity-linked in B_Γ, and

(2) N_Γ, the *domain*, is a nonempty set of individual terms that includes every individual term that occurs in any member of B_Γ, and

(3) S_Γ, the *sentential domain*, is a nonempty set of sentence terms that includes every sentence term that occurs in any member of B_Γ, and

(4) C_Γ, the *context domain*, is a set of context terms, and

(5) f_Γ, the *context assignment function*, is a function whose domain is C_Γ such that for all $\gamma \in C_\Gamma$, $f_\Gamma(\gamma) = \emptyset$.

(The empty set, $\emptyset$, is *not* a context for *AL*.) Now we can define the set of contexts for *AL* stagewise, as follows:

Let M_0 = the set of basic contexts for *AL*, as defined above.

For each $i \geq 0$, let M_{i+1} = the set of quintuples $\langle B_\Gamma, N_\Gamma, S_\Gamma, C_\Gamma, f_\Gamma \rangle$ such that

(1) B_Γ, N_Γ, S_Γ, and C_Γ are as in the definition of basic contexts for *AL* above, and

(2) f_Γ is a function whose domain is C_Γ such that for all $\gamma \in C_\Gamma$, either

(a) $f_\Gamma(\gamma) =$, or

(b) $f_\Gamma(\gamma)$ M_i.

Let the set of contexts for *AL* be $M = \bigcup M_i$, $i \geq 0$.

Supposing that contexts are defined in this way, we may add assertibility and deniability conditions for sentences containing the new vocabulary as follows:

(A**Asst**) If γ C_Γ and $f_\Gamma(\gamma) \neq$ and ϕ is assertible in $f_\Gamma(\gamma)$, then **Asst**([ϕ], γ) is assertible in Γ.

(A**Den**) If γ C_Γ and $f_\Gamma(\gamma) \neq$ and ϕ is deniable in $f_\Gamma(\gamma)$, then **Den**([ϕ], γ) is assertible in Γ.

(A§) If for some context variable ω and some context constant γ, $\phi\gamma/\omega$ is assertible in Γ, then $\S\omega\phi$ is assertible in Γ.

(D**Asst**) If γ C_Γ and $f_\Gamma(\gamma) \neq$ and ϕ is not assertible in $f_\Gamma(\gamma)$, then **Asst**([ϕ], γ) is deniable in Γ.

(D**Den**) If γ C_Γ and $f_\Gamma(\gamma) \neq$ and ϕ is not deniable in $f_\Gamma(\gamma)$, then **Den**([ϕ], γ) is deniable in Γ.

(D§) If for some context variable ω and every context constant γ C_Γ, $\phi\gamma/\omega$ is deniable in Γ, then $\S\omega\phi$ is deniable in Γ.

In light of the new use for sentence terms, we need to revise the deniability conditions for sentences formed with the quantifier Σ, thus:

(DΣ) If for all sentences ψ such that either [ψ] is a member of S_Γ or ψ is assertible or deniable in Γ *or* (this is the new part) for some γ in C_Γ, ψ is assertible or deniable in $f_\Gamma(\gamma)$, $\phi[\psi]/\tau$ is deniable in Γ, then $\Sigma\tau\phi$ is deniable in Γ

Logical validity for *AL* may be defined according to the usual pattern: Where *S* is a finite set of sentences of *AL* and ϕ is a sentence of *AL*, the argument having the

sentences in S as premises and ϕ as conclusion is *logically valid* if and only if for every context Γ for *AL*, if every sentence in S is assertible in Γ, then ϕ is assertible in Γ too.

One thing to notice about these new assertibility conditions is that we do not define the deniability conditions for **Asst**([ϕ], γ) in terms of the deniability of ϕ, which would be the usual pattern; likewise, we do not define the deniability conditions for **Den**([ϕ], γ) in terms of the assertibility of ϕ. Rather, we define the deniability conditions for **Asst**([ϕ], γ) in terms of the *non*assertibility of ϕ, and we define the deniability conditions for **Den**([ϕ], γ) in terms of the *non*deniability of ϕ. The rationale for these departures from the usual pattern is that we want to be able to say in this language that a sentence is neither assertible nor deniable. Given the assertibility conditions laid down in the previous paragraph, a sentence of the form $\neg$ (**Asst**([ϕ], γ) **Den**([ϕ], γ)) will be assertible in Γ provided ϕ is neither assertible nor deniable in $f_\Gamma(\gamma)$. (But (**Asst**([ϕ], γ) **Den**([ϕ], γ)) is not assertible in every context, because it may happen that, for some Γ, γ is not in C_Γ, or γ is in C_Γ but $f_\Gamma(\gamma) =$.)

For an example, let us construct a context in which the sentence Σ**s¶gAsst**(**s**, **g**) ("Some sentence is assertible in every context") is assertible. Suppose:

$B_\Gamma =$, $N_\Gamma = \{$**b**$\}$, $C_\Gamma = \{\Delta, \Lambda\}$, $f_\Gamma(\Delta) = \Psi$, $f_\Gamma(\Lambda) = \Theta$.

$B_\Psi = \{$**Fa**$\}$, $N_\Psi = \{$**a**$\}$, $C_\Psi = \{\Delta\}$, $f_\Psi(\Delta) =$.

$B_\Theta = \{$**Fa**, $\neg$ **Hc**$\}$, $N_\Theta = \{$**a**$\}$, $C_\Theta = \{\Lambda\}$, $f_\Theta(\Lambda) =$.

(The membership of the sentential contexts is irrelevant for purposes of this example.) **Fa** belongs to B_Ψ. So **Fa** is assertible in $\Psi = f_\Gamma(\Delta)$. So **Asst**([**Fa**], Δ) is assertible in Γ. Also, **Fa** belongs to B_Θ. So **Fa** is assertible in $\Theta = f_\Gamma(\Lambda)$. So **Asst**([**Fa**], Λ) is assertible in Γ. But Δ and Λ are the only context terms in C_Γ. So **¶gAsst**([**Fa**], **g**) is assertible in Γ. So Σ**s¶gAsst**(**s**, **g**) is assertible in Γ.

An important feature of the construction of the set of contexts for *AL* is that it can never happen that, for some γ in C_Γ, $f_\Gamma(\gamma) = \Gamma$. That is, for all contexts Γ for *AL*, and for

all γ $\in$ C_Γ, $f_\Gamma(\gamma) \neq \Gamma$. Call this the *well-foundedness assumption* for contexts for *AL*. In the next section, I will defend the reasonableness of this assumption. For now let us observe that in consequence of this feature of the construction of contexts, we cannot formulate in *AL* any paradoxes of assertibility in a context. Here I can only illustrate how the paradoxes are evaded. For a general proof, we would need to prove that in general it cannot happen that for some sentence ϕ and some context Γ for *AL,* both ϕ and $\neg\phi$ are assertible in Γ. More generally, we need to show that, if for each i, $1 \leq i \leq n$, c_i and d_i are identity-linked, then not both $\phi c_1/v_1 \ldots c_n/v_n$ and $\neg\phi\, d_1/v_1 \ldots d_n/v_n$ are assertible in Γ. These things can be done, but I will not take the space to do them here.

For example, suppose α = "α is not assertible in any context". From this it might seem that we could derive a contradiction as follows:

1. α = "α is not assertible in any context".
2. Suppose α is assertible in some arbitrary context Γ.
3. Given 2, "α is not assertible in any context" is assertible in Γ.
4. Given 2, for all c $\in$ C_Γ, "α is not assertible in"$^\frown c$ is assertible in Γ.
5. Given 2, for all c $\in$ C_Γ, "α is assertible in"$^\frown c$ is deniable in Γ.
6. Given 2, for all c $\in$ C_Γ, α is not assertible in $f_\Gamma(c)$.
7. Given 2, α is not assertible in Γ.
8. α is not assertible in any context. (From 2–7.)
9. Suppose α is not assertible in any context.
10. Given 9, "α is not assertible in any context" is not assertible in any context.
11. Given 9, for every context Γ, there is a c $\in$ C_Γ such that "α is not assertible in"$^\frown c$ is not assertible in Γ.
12. Given 9, for every context Γ, there is a c $\in$ C_Γ such that "α is assertible in"$^\frown c$ is not deniable in Γ.
13. Given 9, for every context Γ, there is a c $\in$ C_Γ such that α is assertible in $f_\Gamma(c)$.

14. Given 9, α is assertible in some context.
15. α is assertible in some context. (From 9–14.)
16. α is both assertible in some context and not assertible in any context.

However, there is a fallacy in this argument, in the step from 6 to 7. From the fact that for all c C_Γ, α is not assertible in $f_\Gamma(c)$, it does not follow that α is not assertible in Γ, because we cannot assume that Γ is in the range of f_Γ. Indeed, by the well-foundedness assumption, we can be sure that it is not.

There has to be a fallacy in the second half of the argument too, since we can easily construct a context in which both of the following two sentences are assertible:

α = "α is not assertible in any context".

α is not assertible in any context.

We can even arrange that in that same context, the following sentence is assertible:

"α = 'α is not assertible in any context' " is assertible in every context.

(We took this for granted in the step from 5 to 6 and the step from 12 to 13.) That is, **a = [¬ §gAsst(a, g)]**, **¬ §gAsst(a, g)**, and **¶gAsst([a = [¬ §gAsst(a, g)], g)** may all be assertible in a single context. Such a context Γ may be constructed as follows:

B_Γ = {**a = [¬ §gAsst(a, g)]**}, N_Γ = {**a**}, C_Γ = {Δ}, $f_\Gamma(\Delta) = \Theta$, where:

B_Θ = {**a = [¬ §gAsst(a, g)]**}, N_Θ = {**a**}, C_Θ = {Δ}, $f_\Theta(\Delta) = \Lambda$, where:

B_Λ = {**a = [¬ §gAsst(a, g)]**}, N_Λ = {**a**}, C_Λ = {Δ}, $f_\Lambda(\Delta) = \Omega$, where:

B_Ω = , N_Ω = {**a**}, C_Ω = {Δ}, $f_\Omega(\Delta)$ = .

Clearly, **a = [¬ §gAsst(a, g)]** is assertible in Γ. Since it is also assertible in $f_\Gamma(\Delta) = \Theta$, and Δ is the only member of C_Γ, **¶gAsst([a = [¬ §gAsst(a, g)]], g)** is assertible in Γ. To show that **¬ §gAsst(a, g)** is assertible in Γ, we have to do a little more work: Observe

that **Asst**(**a**, Δ) is not deniable in Ω; so §**gAsst**(**a**, **g**) is not deniable in Ω; so ¬ §**gAsst**(**a**, **g**) is not assertible in Ω = f_{Λ}(Δ). So **Asst**([¬ §**gAsst**(**a**, **g**)], Δ) is deniable in Λ. Since **a =** [¬ §**gAsst**(**a**, **g**)] is assertible in Λ, **Asst**(**a**, Δ) is deniable in Λ. Since Δ is the only context term in C_{Λ}, for all γ in C_{Λ}, **Asst**(**a**, γ) is deniable in Λ. So §**gAsst**(**a**, **g**) is deniable in Λ. So ¬ §**gAsst**(**a**, **g**) is assertible in Λ = f_{Θ}(Δ). So **Asst**([¬ §**gAsst**(**a**, **g**)], Δ) is not deniable in Θ (indeed, it is assertible in Θ). Since **a =** [¬ §**gAsst**(**a**, **g**)] is assertible in Θ, **Asst**(**a**, Δ) is not deniable in Θ (it is assertible). So for some γ in C_{Θ}, **Asst**(**a**, γ) is not deniable in Θ. So §**gAsst**(**a**, **g**) is not deniable in Θ (but is assertible). So ¬ §**gAsst**(**a**, **g**) is not assertible in Θ = f_{Γ}(Δ). So **Asst**([¬ §**gAsst**(**a**, **g**)], Δ) is deniable in Γ. Since **a =** [¬ §**gAsst**(**a**, **g**)] is assertible in Γ, **Asst**(**a**, Δ) is deniable in Γ. So for all γ in C_{Γ}, **Asst**(**a**, γ) is deniable in Γ. So §**gAsst**(**a**, **g**) is deniable in Γ. So, finally, ¬ §**gAsst**(**a**, **g**) is assertible in Γ. The precise location of the fallacy in the second half is the step from 13 to 14. As the example shows, from the fact that for every context (consider Θ) that we can talk about in our context (which might be Γ), there is, from the point of view of that context (Θ) a context in which α is assertible (consider Λ), it does not follow that from the point of view of our own context (Γ) there a context in which α is assertible (α not being assertible in Θ).

For a simpler example of how paradox is avoided, suppose we had a context term "Δ" such that f_{Γ}("Δ") = Γ and a sentence β, which read "β is not assertible in Δ". Then it might appear that from the fact that β = "β is not assertible in Δ" we could derive the contradictory conclusion that β both is and is not assertible in Γ (see my 2003, p. 209). However, the derivation will not go through without the assumption that f_{Γ}("Δ") = Γ, and that equation is precluded by the well-foundedness assumption for contexts for *AL*.

8. The context we are in

The language *AL* is consistent in the sense that there is no sentence ϕ of *AL* such that both ϕ and ¬ϕ are assertible in a single context. Furthermore, if for each i, $1 \leq i \leq n$, c_i and d_i are identity-linked in Γ, then not both $\phi c_1/v_1 \ldots c_n/v_n$ and $\neg\phi\, d_1/v_1 \ldots d_n/v_n$ are assertible in Γ. (These claims require a proof, which I am not giving here.) It is essential to these results that the context assignment function for a context Γ cannot assign Γ itself to any context term in the context domain for Γ. In other words, we must deny the possibility of anyone's talking about the context that he or she is in. This is what I have called the well-foundedness assumption. Only due to this limitation are we able to block the derivation of contradictions from plain facts such as the fact that α = "α is not assertible in any context". That restriction is guaranteed by our construction of the set of contexts. But now we must ask whether the well-foundedness assumption is a reasonable restriction that we can motivate independently.

As I explained in section 5 above, we are to think of the context pertinent to a conversation as comprising everything that is relevant to the conversation in light of the goals of the conversation and the actual circumstances in which it takes place. The context for a conversation, so conceived, is objective in the sense that the participants in the conversation may be entirely mistaken about the contents of the context. What is relevant to the goals of the conversation in light of the actual external circumstances does not depend on the interlocutors' states of mind.

In light of this conception of the context as comprising what is objectively relevant to the conversation, we can perhaps understand the necessity of the well-foundedness assumption. Suppose that Γ is the context *pertinent* to a conversation *C*, that is, the context relative to which the assertibility of sentences in *C* ought to be evaluated. Then the content of Γ is a matter of what is relevant to *C*. In particular, a context Δ is in the range of the context assignment function for Γ only if Δ is relevant to

C. Thus, to defend an account of the content of context Γ would be to establish the relevance to *C* of every context that is taken to belong to the context assignment function for Γ. So if Γ itself were in the range of the context assignment function for Γ, then we would be in the impossible position of having to establish the relevance of Γ to *C* before we had established the content of Γ. So, assuming that it will be possible to defend an account of the content of Γ, Γ cannot itself be a member of the range of the context assignment function for Γ. It is fair to assume that it must be possible to defend an account of the content of a context, because, while contexts may be objective, they must also be the sort of thing whose content can in principle be discovered.

It may seem to us on occasion as though we were referring to the very context we are in. For instance, if we are madly driving around desperately trying to find the location where we are supposed to return the rented car we are driving, you might say to me, "Last time, you didn't have any problem", and I might reply, with irritation, "That's not relevant in this context; it does not help us one bit". Here it might look as though I was saying something assertible about what was assertible in the context we are in. But it is not actually so obvious that the context in which my sentence is assertible is the same as the context I am talking about. We can carry on multiple conversations that interweave with one another in time, and it might very well be that when I utter my sentence I am, so to speak, taking a break from the conversation we had been having, whose goal was to find the rental car return office, in order to have a brief conversation about that conversation, where the goal of this second conversation is to correct your manner of contributing to the first.

This limitation on what we can talk about is at the same time a limitation on our ability to formulate the semantics for our own language. Certainly, we may formulate various generalizations about assertibility in a context. We might, for instance, declare an argument valid, meaning that for every context in which the premises are assertible, the conclusion is assertible as well. But the context pertinent to the conversation in which we

assert this generalization is not in the range of the context assignment function for that context, and so, as we can observe from the perspective of another context, our generalization is not really about absolutely every context. This limitation need not matter in any practical way at all; for it may still be the case that, in any context in which the question arises, the same generalization will be assertible; in no context is that form of words deniable. Of course, even in saying this I am failing to speak of absolutely every context.

Many philosophers, from diverse traditions, have perceived an analogous quandary pertaining to reference to one's self. Consider, for instance, this passage from Kant:

> Through this I or he or it (the thing) that thinks nothing more is represented than a transcendental subject of thoughts = *x*, which cognizes itself only through the thoughts which are its predicates and of which, separated out, we can never have the slightest concept; consequently we perpetually revolve in a circle around it, in that we must always make use of its representation in order to form some judgment of it. (Kant 1956/1781, A346/B404)

A similar thought forces itself on Wittgenstein:

> If I wrote a book called *The World as I Found It*, I should have to include a report on my body, and should have to say which parts were subordinate to my will, and which were not, etc., this being a method of isolating the subject, or rather of showing that in an important sense there is no subject; for it alone could *not* be mentioned in that book. (Wittgenstein 1961/1921, 5.631)

Here is Jean-Paul Sartre on the same theme:

> Thus consciousness (of) belief and belief are one and the same being, the characteristic of which is absolute immanence. But as soon as we wish to grasp this being, it slips between our fingers, and we find ourselves faced with a pattern of duality, with a game of reflections. For consciousness is a reflection (*reflet*), but *qua* reflection it is

> exactly the one reflecting (*réfléchissant*), and if we attempt to grasp it as reflecting, it vanishes and we fall back on the reflection. (Sartre 1956/1943, pp. 75-76)

The idea is expressed most plainly, with the least implication of profound mystery, by Gilbert Ryle:

> A higher order action cannot be the action upon which it is performed. So my commentary on my performances must always be silent about one performance, namely itself, and this performance can be the target only of another commentary. Self-commentary, self-ridicule and self-admonition are logically condemned to eternal penultimacy. Yet nothing that is left out of any particular commentary or admonition is privileged thereby to escape comment or admonition for ever. On the contrary it may be the target of the very next comment or rebuke. (1949, p. 195)

What Kant, Wittgenstein, Sartre and Ryle are all remarking upon is a difficulty in attempting to reflect on one's own self. Necessarily some aspect of oneself is omitted from the content of one's reflection, namely, that very act of reflection.

We do not have to read these passages as denying the possibility of self-reference. These philosophers do not claim that a thought cannot refer to itself, although their claims may entail that there will always be something about itself that such a thought fails to represent. Moreover, these philosophers cannot be understood as imposing a hierarchy. They are not suggesting that the self is sliced into moments or planes and that one's reflection on oneself always belongs to a different plane from the self reflected upon. Their claim is rather that, for any given representation that one may form of oneself, there is always an aspect of oneself that falls not within the purview of that representation, namely, that aspect of oneself that consists in one's representation of oneself.

As these philosophers claim that an act of reflection always fails to encompass itself, I claim that a generalization over contexts always fails to encompass the context with respect to which the generalization ought to be judged assertible or not. That is an analogous point that remains after we have abandoned referential semantics in favor of

contextual semantics and have taken the primary locus of conceptual representation to be the assertions of interlocutors in conversation rather than the judgements of isolated minds. Likewise, just as these philosophers supposed that any act of reflection can become the object of a further reflection, so too we allow that any context of assertion can fall within the scope of generalizations assertible in other contexts. One's sense that one can in fact talk about the context one is in stems from the knowledge that one can always shift to a different context and from that point of view talk about the context one was in just before.

References

Beall. JC. 2001. "A Neglected Deflationist Approach to the Liar", *Analysis*, 61, 126-129.

Beall, JC. 2002. "Deflationism and Gaps: Untying 'Not's in the Debate," *Analysis*, 62, 299-305.

Beall, JC and Bradley Armour-Garb. 2003. "Should Deflationists be Dialethists?", *Noûs*, 37, 303-324.

Cartwright, Richard, 1994. "Speaking of Everything"', *Noûs* 98: 1-20.

Etchemendy, John, 1990. *The Concept of Logical Consequence*, Harvard University Press.

Field, Hartry. 1994*a*. "Deflationist Views of Meaning and Content", *Mind*, 103, 249-285. (Reprinted in Field 2001.)

Field, Hartry. 1994*b*. "Disquotational Truth and Factually Defective Discourse", *Philosophical Review*, 103, 405-452. (Reprinted in Field 2001.)

Field, Hartry. 2001. *Truth and the Absence of Fact*, Oxford University Press.

Forster, T. E. 1992. *Set Theory with a Universal Set: Exploring an Untyped Universe*, Oxford University Press.

Gauker, Christopher. 1997. "Universal Instantiation: A Study of the Role of Context in Logic", *Erkenntnis*, 46, 185-214.

Gauker, Christopher. 1998. "What is a Context of Utterance?", *Philosophical Studies*, 91, 149-172.

Gauker, Christopher. 1999. "Logic and Deflationism", *Facta Philosophica*, 1, 167-196.

Gauker, Christopher. 2001. "T-Schema Deflationism versus Gödel's First Incompleteness Theorem", *Analysis*, 61, 129-136.

Gauker, Christopher. 2003. *Words without Meaning*, MIT Press.

Gauker, Christopher. forthcoming. *Conditionals in Context*, MIT Press.

Holton, Richard. 2000. "Minimalism and Truth-Value Gaps", *Philosophical Studies* 97: 137-168.

Horwich, Paul. 1990. *Truth*, Blackwell.

Kant, Immanuel. 1956/1781. *Kritik der Reinen Vernunft*, ed. Raymund Schmidt, Felix Meiner Verlag.

Kaplan, David. 1989. "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals", in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan*, Oxford University, pp. 481-564.

Kreisel, Georg, 1967. "Informal Rigour and Completeness Proof", in Imre Lakatos, ed., *Problems in the Philosophy of Mathematics*, North-Holland, pp. 138-157.

Kripke, Saul. 1975. "Outline of a Theory of Truth", *Journal of Philosophy*, 72, 690-716.

Leblanc, Hugues. 1976. *Truth-Value Semantics*, North-Holland.

McGee, Vann. 1992. "Maximal Consistent Sets of Instances of Tarski's Schema (T)", *Journal of Philosophical Logic*, 21, 235-241.

Rayo, Agustín and Timothy Williamson. 2003. "Unrestricted First-Order Languages", in JC Beall, ed., *Liars and Heaps: New Essays on Paradox*, Oxford University Press, pp. 331-356.

Ryle, Gilbert. 1949. *The Concept of Mind*, Barnes and Noble Books.

Sartre, Jean-Paul. 1956/1943. *Being and Nothingness*, tr. Hazel E. Barnes, Philosophical Library.

Simmons, Keith. 1993. *Universality and the Liar: An Essay on Truth and the Diagonal Argument*, Cambridge University Press.

Stalnaker, Robert. 1974. "Pragmatic Presuppositions", in Milton K. Munitz and Peter K. Unger, eds., *Semantics and Philosophy*, New York University Press, pp. 197-213

Wittgenstein, Ludwig. 1961/1921. *Tractatus logico-philosophicus*, trans. D. F. Pears and B. F. McGuiness, Routledge and Kegan Paul.