NUEL BELNAP

ON RIGOROUS DEFINITIONS

(Received 27 February 1993)

Definitions are crucial for every serious discipline.[1] Here I consider them only in the sense of explanations of the meanings of words or other bits of language. (I use "explanation" as a word from common speech, with no philosophical encumbrances.) As a further limitation I consider definitions only in terms of well-understood forms of rigor. Prominent on the agenda will be the two standard "criteria" — eliminability and conservativeness — and the standard "rules". There is, alas, hardly any literature on this topic. The discussion will therefore be preliminary, all too elementary, and imperfectly plain.

1. SOME PURPOSES OF DEFINITIONS

There are two especially clear social circumstances that call for a meaning-explaining definition, and then many that are not so clear. The clear ones call either for (1) "dictionary" definitions or for (2) "stipulative" definitions; one of the less clear circumstances calls for an (3) "analysis."

1.1. *"Dictionary" or "Lexical" Definitions*

One might need to explain the existing meaning of an *old* word; that is, a word already in use in the community, but unfamiliar to the person wanting the explanation.

1. EXAMPLE. (Lexical definitions) (1) What is a *sibling*? A sibling is a brother or a sister. (2) What does it mean to *square* a number? One obtains the square of a number of multiplying it by itself. (3) What do you mean by *zero*? Zero is the least integer. (4) What is a *brother*? A brother is a male sibling.

Philosophical Studies **72**: 115—146, 1993.

I postpone discussion of (a) the question whether these interchanges should be sprinkled with quotation marks and (b) the circularity threatened by combining the sibling/brother examples.

1.2. *"Stipulative" Definitions*

One might wish to explain a proposed meaning for a *new* word. The purpose might be to enrich the language by making clear to the community of users that one intends that the new word be used in accord with the proposed meaning. This case squarely includes the putting to work of an "old" word with a new technical meaning.

2. EXAMPLE. (Stipulative definitions) (1) Let *omega* be the first ordinal after all the finite ordinals. (2) Let a *group* be a set closed under a binary operation satisfying the following principles . . . . (3) By a *terminological realist* I mean a philosopher who subscribes to the following doctrines . . .

The person who introduces a new word might have any one of various purposes. Here is a pair: the person might want a mere abbreviation, to avoid lengthy repetition ("By *E2* I refer to volume 2 of *Entailment: The Logic of Relevance and Necessity*"); or the person might take it that he or she is cutting at a conceptual joint (a *complete lattice* is a partially ordered set in which every subset has a least upper and a greatest lower bound).

1.3. *"Analyses" or "Explications"*

There are many cases not exhibiting either of these clear purposes, including perhaps most distinctively philosophical acts of definition; in these cases (Carnap calls some of them "explications") one wants both to rely on an *old*, existing meaning and to attach a *new*, proposed meaning; it seems that one's philosophical purposes would not be served if one let go of either pole.

3. EXAMPLE. (Explicative definitions) (1) Let *knowledge* (in the present technical sense) be justified true belief. (2) We say that *A implies B* if *B* is true on every interpretation on which *A* is true.

Observe that in these cases the philosopher neither intends simply to be reporting the existing usage of the community, nor would his or her purposes be satisfied by substituting some brand new word. In some of these cases it would seem that the philosopher's effort to explain the meaning of a word amounts to a proposal for "a good thing to mean by" the word. I learned this phrase from Alan Ross Anderson. Part of the implication is that judging philosophical analyses is like judging eggs: There are no shortcuts; each has to be held to the candle.

1.4. *Invariance of Standard Theory Across Purposes: Criteria and Rules*

The extraordinary thing is this: The applicability of the standard theory of definition remains invariant across these purposes — just so long as the purpose is to "explain the meaning of a word." (The phrase "a word" suggests that it is a single word that is being defined, whereas the standard theory can certainly do better than that. For simplicity, however, I omit discussion of the simultaneous explanation of several terms by means of a single act of definition.)

This standard theory has two parts. In the first place, it offers two *criteria* for good definitions: the criterion of eliminability (which requires that the defined term be eliminable in favor of previously understood terms) and the criterion of conservativeness (which requires that the definition not only not lead to inconsistency, but not lead to anything — not involving the defined term — that was not obtainable before). In the second place, it offers some *rules* for good definitions, rules which if followed will guarantee that the criteria are satisfied.

History. The standard theory of definitions seems to be due to Leśniewski, who modeled his "directives" on the work of Frege, but I cannot tell you where to find a history of its development. The standard citation seems to be Leśniewski 1931; see also Leśniewski 1981 (*Collected works*). I learned most of the theory first from Suppes 1957, who credits Leśniewski (p. 153, note). There should have been mini-histories in either Church 1956 or Curry 1963, but I couldn't find what I was looking for. The matter was well understood by Frege (e.g. in Frege 1964), Couturat (see Couturat 1905), Carnap (e.g. in Carnap 1937) and Tarski (see e.g. Tarski 1941 for some well-chosen elemen-

tary words). Tarski himself contributed heavily to the theory, as evidenced in the material translated and reprinted in Tarski 1956. There Tarski gives the dates and circumstances of his own early contributions in the 20s and 30s. But no one of these lays out an account of the history of the matter in its beginnings. The standard histories of logic (Bochenski 1956, Kneale and Kneale 1962) do not discuss modern theories of definition. Neither does Kneebone 1963. Neither does Church's article on "definition" in Runes 1962. The 207-page book Robinson 1950 neither discusses the technical theory nor refers to its history (though there is some reference to the history of nontechnical discussions). The definition article in *The Encyclopedia of Philosophy* (1967) does not even mention Leśniewski. The only useful general references I happen to know are the definition article in the *Dictionary of Logic*, Marciszewski 1981, and some penetrating paragraphs and authoritative citations in Luschei 1962 (see especially pp. 36—37 and nn. 34 and 78).

The *criteria* are like the logician's account of semantic consequence: The latter is proposed as an account of "good inference," the former as an account of "good definition." And the standard *rules* are like the logician's rules defining derivability in a particular system: If you follow the logician's rules for constructing derivations in his system, you will derive all and only correct semantic consequences, i.e., you will make all and only "good inferences." In the same way, if you follow the logician's rules for constructing definitions, you will offer all and only definitions that satisfy the criteria of eliminability and conservativeness. You will, that is, offer all and only "good definitions." The analogy runs deep: Just as it is a hard theorem that in first order logic semantic consequences and derivability (in some one formal system) are in agreement, so it is a hard theorem that the two criteria and the standard rules are in agreement. The hard part is Beth's Definability Theorem. There is an additional parallel. We know that agreement between semantic consequence and derivability (in some one formal system) gives out when first order logic is enriched in any one of several respects. So does the agreement between the criteria and the rules of definition.

I discuss the two criteria in some length, and then say just a little about the rules.

2. ELIMINABILITY AND CONSERVATIVENESS I

The standard criteria for good definitions are those of "eliminability" and "conservativeness." Where do these criteria come from? Why should we pay attention to them? Indeed, we may and shall raise the presupposed question of whether we *should* pay them attention. It is surely remarkable that philosophical-logical literature is nearly silent on these questions.

I have only the most elementary of answers to propose. It seems to me that we can derive a motivation for these criteria from the concept of definitions as explanations of the meanings of words (or other bits of language). Under the concept of a definition as explanatory, (1) a definition of a word should explain *all* the meaning that a word has, and (2) it should do *only* this and nothing more. That a definition should (1) explain *all* the meaning of a word leads to the criterion of eliminability. That a definition should (2) *only* explain the meaning of the word leads to the criterion of conservativeness. Observe that (1) and (2) are not quite analogous in their respective placements of "all" and "only." This warns us that, as we shall see, the two criteria are also not exactly analogous.

2.1. *Criterion of Eliminability (Rough Account)*

One may approach the criterion of eliminability from the direction of the "use" picture of meaning, with picture-slogan, "meaning is use." Then to explain *all* its uses, that is, its use in *every* context. (I intend that this recipe neglect ambiguity). The advance is this: The metaphorical quantifier in "*all* the meaning" is cashed out in terms of a nonmetaphorical (if still imprecise) quantifier over contexts.

And what is it to explain the meaning of a word in a single context? That depends on what counts as a context. The tradition has found it useful to concentrate on *(declarative) sentential contexts.* I shall do so as well, even though these contexts are not enough for an adequate story about language. So what is it to explain the meaning of a word in a single sentential context, say B? The standard move, and a good one, is to say that it is to explain the meaning of the containing sentence B. The next standard idea is this: To explain the meaning of a sentence is

to explain its *role in inference*. Combining ideas yields the doctrine that to explain all the meaning of a word is to explain the inferential connections of each containing sentence *B*.

At this stage there enter two new elements. First, explanations quite generally are more prized if they are given in terms that are previously understood. This we may call *noncircularity*. Second, in favorable situations we can hope to explain the inferential role of a new sentence by identifying its role with that of an old sentence. This we may call *no inferential enrichment*. (I apologize for the jargon.) The combination of these two elements comes to this: There is in the language another sentence that (a) does not contain the defined term and (b) occupies exactly the same inferential role as that wanted for the containing sentence *B*. The standard account presupposes that the situation is favorable. Accordingly, the criterion of eliminability requires that for *each* sentential context *B* containing the defined word, the definition give enough information to allow formation of an inferentially equivalent piece of language *B′* that contains only previously understood words. Then, and only then, will we be sure that we have explained *all* the meaning of the word to be defined — whether our purpose is to explain an existing meaning of an *old* word or to give a new meaning to a *new* word. In either case, for each context in which the defined word can grammatically occur, we must provide enough information to allow construction of an equivalent statement that contains only previously understood words. Under the assumption that the defined word is not "previously understood," what is then required is enough information to permit "elimination" of the defined word in favor of other words that are previously understood.

The above account needs considerable refinement; containing as it does several obscure phrases, it cannot itself count as a definition of "satisfies the criterion of eliminability" that satisfies the criterion of eliminability. I postpone this necessary refinement until after a preliminary account of the criterion of conservativeness, but pause to illustrate some of the obscurity with an example.

4. EXAMPLE. (Holiness) Holiness is what the gods all love (Euthyphro).

Taken as a definition, and read in a standard two-valued way, this

permits elimination of "holy" in "extensional" contexts, which is good. As Socrates pointed out, however, it does not permit elimination of "holy" in "becausal" contexts. Socrates asked Euthyphro to consider the following plausible premiss.

> Holy acts are loved by the gods because they are holy.

If we tried to use Euthyphro's Example 4 to eliminate the second occurrence of "holy" — the one in the scope of "because" — we should find that we were led to the following implausible conclusion.

> Holy acts are loved by the gods because they are loved by the gods

This leads us to see that "context" needs specification.

We should keep in mind quotation contexts, and also psychological contexts involving rapt attention such as "he was turning over in his mind the question of whether prosecution of his father was holy." If we do then we will be more likely to keep remembering that *no* definition will permit elimination in absolutely *every* context. Such remembering will perhaps disincline us to ask more of a definition than it can very well deliver. But perhaps not. Some people respond by cooking up an account of quotes (for example) on which they do not count as "contexts." This seems to me an unhelpful obfuscation of the general principle.

The matter of contexts is easily overlooked and critically important. Since no definition can sensibly license universal elimination, any sensible act of definition must include a reference to the family of intended contexts. The point is easily overlooked because those who care about the proper use of definitions frequently think of themselves as already having specified some background family of contexts, e.g. the first order functional calculus.

The matter of no inferential enrichment is also easily forgettable, but should not be forgotten. It is easy to describe languages in which there are definite inferential roles whose status is changed from "unoccupied" to "occupied" by stipulative definitions exploiting precisely those roles. Here is an uncomplicated example.

5. EXAMPLE. (The Absurd) Let the language be entirely

> positive, lacking any trace of negation, but be otherwise standard. Stipulate that to say of anything that it is *F* is to say something that implies everything. Thus each sentence *Fa* pays the role of The Absurd.

This evidently defines an inferential role for each sentence of which each *Fa* is a part, and thus explains all the meaning of *F*. But since no sentence already in the language implies every sentence, the definition is inferentially enriching and does not satisfy eliminability. (Observe that the failure of eliminability is *not* due to a violation of noncircularity.)

2.2. *Criterion of Conservativeness (Rough Account)*

On the standard view a good definition (in the sense of an explanation of meaning) should not only explain *all* the meaning of the word, as required by the criterion of eliminability. It should *only* do this. It should be empty of assertional content beyond its ability to explain meaning. If it were to go beyond the job assigned, say by claiming that Euthyphro is pious, it might indeed be doing something useful. It would, however, no longer be entirely satisfactory as a *definition*. In this perfectly good sense, a definition should be neither true nor false (whether explaining an *old* word or introducing a *new* one): Whether or not it has a truth value qua definition (we might argue about that), it should make no claims beyond its explanation of meaning.

There is a special case to which I shall not return. Clearly a definition should not take you from consistency to inconsistency; that would be a dreadful violation of conservativeness. The older theoreticians of definitions in mathematics were, however, insufficiently severe when they suggested that consistency is the *only* requirement on a mathematical definition.

Terminology for this criterion is a little confusing. Some folks use "conservative" to mean something like "does not permit the proof of anything we couldn't prove before." This accounts for calling the second criterion that of conservativeness: A definition satisfying the second criterion is conservative in that very sense. Other folks use "creative" to mean something like "permits proof of something we

couldn't prove before." It is with this sense in mind that the second criterion is sometimes called "the criterion of noncreativity." I opt for the former name merely to avoid the negative particle. Under either name, the criterion demands that the definition not have any consequences (other than those consequences involving the defined word itself) that were not obtainable already without the definition. If before the definition we could not establish that Euthyphro is pious, and if the definition is of neither "Euthyphro" nor "pious" (but perhaps of "zero" or "sibling"), then it should not be possible to use the definition to show that Euthyphro is pious. Were we able to do so, the definition would manifestly contain more information than a mere explanation of meaning: It would not be conservative. Neither in explaining the meanings of old words nor in introducing new words should one use the cover of definitions to smuggle in fresh assertions. It's bad manners.

6. EXAMPLE. (Noid) Suppose that Wordsmith introduced "noid" by announcing that all the guys on the other side of Divider Street are noids (a sufficient condition) and that all noids are born losers (a necessary condition).

Wordsmith's plan appears innocent: He wishes to help his gang understand "noid" by squeezing it between two understood ideas. Wordsmith's introduction of "noid," however, may or may not be innocent. If the gang and Wordsmith have already committed themselves to the proposition that all guys on the other side of Divider Street are born losers, then Wordsmith has not by this "definition" violated conservativeness. But if Wordsmith has not done this, if he has not committed himself and the gang to the born-loserness of all guys across Divider, then he has violated conservativeness.

This consideration is easily mixed with two others. There is in Example 6 an obvious difference in the emotional and practical freighting of the two bits connected via "noid." The one starts out in the unloaded "space language," as Carnap might say, while the other suggests contempt and a practice of intimidation. What should not be mixed is the division between sufficient and necessary conditions on the one hand, and the division between unfreighted and freighted language on the other. Furthermore, neither should be mixed up with what is related to the possibility of justificatory warrant.

There is a middle case. Suppose Wordsmith is participating in a practice according to which his announcement *combines* (1) the assertion that all guys across Divider Street are born losers with (2) the conservative introduction of "noid." There are two crucially different subcases. *Revealed assertion*: The practice is such that Wordsmith's form of words bears on its surface the division into assertion and introduction. *Concealed assertion*: The practice is such that the form of words conceals the assertional content or even misleads into the false supposition that there isn't any. I shall pursue the difference between revealed and concealed assertion below.

7. EXAMPLE. (Empty set) Suppose Janet says "Definition: ∅ is that set that has no members."

If Janet's community has *already* fixed on a theory that implies that *there is* a unique set with no members, she is all right. But since this "definition" permits her to prove that there is a set with no members (by existential generalization), she violates conservativeness if she couldn't prove it before. She would then be using the "definition" to smuggle in an existence claim to which she is not entitled.

Suppose the worst, however, that Janet's community has a Leśniewski-like theory according to which no set has no members. Suppose she goes ahead and introduced ∅ in accord with Example 7. What should we say about her? What did she denote by ∅? Probably no one cares, but there are cases like Janet's about which philosophers do seem to care. For example Leverrier introduced "Vulcan" as the planet responsible for certain astronomical anomalies. Later it was learned that relativistic considerations sufficed for the anomalies; there was no such planet. What shall we then say about the meaning of "Vulcan" and of all the discourse using this term? My own view is this. You should say that the answer to "Is there a unique planet responsible for the astronomical anomalies at issue?" is "No." You should say that the answer to "Is the definition defective?" is "Yes." You should say that some of what we prize in science and other activities does not rest on answers to these eminently clear questions, since we often manage to get along very well when presuppositions fail. You should add that nevertheless *there are no general policies for what to do or what to say in the presence of defective definitions.* Just about anything you remark beyond this will be

either artificial or unclear, and is likely only to add to the general babble of the universe.

It is easy to imagine languages that allow nonconservative or creative definitions in the following sense: In them a single speech act both introduces the new terminology and makes the assertion needed to justify that introduction. In fact English, even scientific English, is like that. Our practices are lax. I suppose that is one thing people mean when they advise us to abandon the analytic-synthetic distinction.

We should not, however, follow this advice. Those philosophers who wish to be clear should follow Carnap in maintaining the analytic-synthetic distinction to the maximum extent possible. That is, when confronted with a complex speech act such as that of Leverrier, it is good to idealize it as consisting of two distinct components: the assertion of the existence and uniqueness of an anomaly-causing planet, and the definition of Vulcan as this very planet. There is, I think, no other way to continue with a *clear* conversation about the matter. For example, I do not think we can otherwise make a clear distinction between revealed and concealed assertions. And we should do so. Speech acts that conceal assertions are (whatever their intentions) misleading. We should try to avoid them whenever we are clear enough to be able to do so. This, as we may infer from the history of even the best science, may not be often. Even when we are muddled, however, we should at least try not to pretend to more clarity than the sad facts warrant.

2.3. *Joint Sufficiency*

In conclusion of this preliminary account, I add the following. The proposal is that each of the criteria of eliminability and conservativeness are necessary for a good definition in the sense of an explanation of meaning; and that together they are sufficient (so that their conjunction provides a good definition of "good definition"). The pros and cons of necessity are easily discussed through examples. In contrast, I do not much know how to defend the claim to sufficiency other than by a rhetorical question: You've asked for a good definition of this word, and now I've given you a procedure for eliminating it from every intended context in favor of something you yourself grant is entirely

equivalent; and I've done nothing else (i.e., I haven't slipped any further statements past you); so what more do you want? Until this rhetorical device can be replaced by something better, there is conspicuous work to be done.

3. ELIMINABILITY AND CONSERVATIVENESS II

Such a rough account of the criteria of eliminability and conservativeness calls for an increase in precision. The standard account provides exactly this.

3.1. *Limits of Applicability*

You cannot buy an increase in precision without paying in the good coin of limitation. Here the costs are of three kinds.

First, the discussion applies only to a community of language users whose language can profitably be described by labeling it an "applied first order language." You should imagine that this language is a fragment of English, a fragment that the users think of as structured in terms of predicate, function, individual, and perhaps sentence constants; truth-functional connectives; individual variables; quantifiers; and identity. But you should *also* imagine that this language is *really used* and therefore has lots of English in it. Think of the language that some mathematicians sometimes employ: a mixture of English and notation, but with the "logic" of the matter decided by the first order calculus. For example, the language will use English common nouns such as "set"; but only in locutions such as "Some sets are finite" that the users think of as translatable in the usual first order way.

I need to say more about what this first limitation means. So far I have said that the users take the *grammar* of their language in the standard first order way. They also think of their *proof theoretical* notions as given in the standard way: I am thinking of "axioms," "rules," "theoremhood," "derivability from premisses," "theory," "equivalent relative to a theory," and so forth. And lastly, they think of their *semantic* concepts in the standard way: "logical truth," "(semantic) consequence of some premisses," "(semantic) equivalence relative to a theory," and so forth. They know, through Gödel, that there is agree-

ment between the appropriate proof-theoretical and semantic ideas. I shall refer to these matters by saying that the first order language *has an inferential use*, or by saying that its sentences occur in *inferential contexts.*

In the second place, I consider only definitions (explanations of meaning) of predicate constants and function constants and individual constants. This limitation is made reasonable (but is certainly not forced) by the previous decision to consider only a first order language. There is no consideration, for example, of Russell's treatment of definite descriptions, or of the meaning of a convention involving the dropping of outermost parentheses, or of quantifiers of a certain style having some limited range.

In the third place, I am going to follow the standard account in considering only definitions that are themselves sentential, and that are in the very same language containing the defined and defining terms. In imposing this limitation I do not intend to be deciding whether the act of defining is "really" imperatival instead of assertional, or "really" metalinguistic rather than not. I do intend to assert, however, that the technical theory of definition goes very much more easily if in giving it one has to talk about only one language: the language including the defined term, the defining terms, and the definition itself. In fact we do need to consider two languages with respect to vocabulary: the language with, and the language without the newly defined term; and that is quite enough in the direction of complication for such an elementary discussion as this.

We can therefore see that the policy, exhibited in Example 1, of giving example definitions above in the "material mode" (as Carnap puts it) correctly forecasts this decision. By taking the definition of "sibling" to be "Anything is a sibling if and only if it is a brother or a sister" instead of "Replace 'sibling' by 'brother or sister' wherever found!", the technical work is greatly simplified, as you will see for yourself if you try to spell out a technically adequate theory of definitions of (say) the metalinguistic imperative kind.

But, you may say, isn't a definition a speech act? How can you pretend adequacy for a theory of definitions that does not contain a theory of such acts? Answer: There is no pretense. The standard account is not adequate to the aim you have envisaged. There is,

however, no existing rigorous and helpful explicit theory of definitions in the guise of speech acts carried out by agents. Such a theory would need to be founded on a general theory of speech acts carried out by agents. And this in turn would need to be founded on a general theory of agency. Perhaps the Belnap/Perloff "seeing to it that" approximation to agency ("stit theory") is sufficient for the purpose; it seems plausible that this should be so. But the work remains to be done.

3.2. *The Terms of the Standard Account*

Recall that the focus is on definitions (1) that are of individual constants, predicate constants, and function constants; (2) that are in the "material mode"; and (3) that are in (not an arbitrary language but) an applied first order language with an inferential use. This focus permits us to give the criteria of eliminability and of conservativeness themselves in the form of definitions, namely, definitions of "satisfying the criterion of eliminability" and of "satisfying the criterion of conservativeness."

It turns out that there are four key entities involved in explaining the standard ideas of eliminability and conservativeness; or, as one might well say, each of these ideas is explained as a four-place predicate. I shall attach to each of these key entities a variable, but for intelligibility I shall use for each a spelled-out variable having a certain mnemonic force.

- *Theory* is the background *theory* (before the definition) in the context of which the definition is being entered. *Theory* is a set of sentences of the usual kind; sometimes it is best to think of it as closed under semantic or proof-theoretic consequence, but my use of the variable is such that often *Theory* can be coded as a set of axioms. If the background theory is thoroughly interpreted, *Theory* can even be the set of truths.
- *Previous-definitions* is the set of *previous definitions* — a set of sentences. This element is required to permit discussion of circularity (a brother is a male sibling; a sibling is a brother or sister).

- *Definition* will be the *definition* itself. I remind you that *Definition* is a sentence on a par with the rest of the sentences in the context of which the act of definition is taking place. One can therefore meaningfully and without quibble speak of its grammatical and deductive and semantic properties and relationships. It is just another first order sentence.
- *Symbol* is the new *symbol* being defined. Recall that *Symbol* is to be a (new) predicate, operator, individual constant, or perhaps a sentence constant. The deepest theory would not require or even permit *Symbol* to be a symbol, permitting it or even requiring it instead to be a grammatical function of an appropriate type; but I shall not in this discussion plumb that particular abyss.

The above discussion has mixed intuitive and rigorous considerations. Here I specify the latter, in particular suppressing all verbs of psychological attitude.

- *Theory* is a set of sentences.
- *Previous-definitions* is a set of sentences.
- *Definition* is a sentence.
- *Symbol* is a symbol (a predicate, operator, individual constant, or perhaps sentence constant).

3.3. *Criterion of Eliminability (Rigorous Account)*

It is important to have the standard account of eliminability made rigorous so that one can more plainly see what is at stake. As promised, the account is presented as a definition of a four-place predicate. Recall that it is a presupposition of the discussion that sentences are made exclusively by way of truth functions and quantifiers.

8. DEFINITION. (Eliminability) Let *Theory*, *Previous-definitions*, *Definition*, and *Symbol* be just as above. Then *Definition*, qua definition of *Symbol*, satisfies the *criterion of eliminability* relative to *Theory* and *Previous-definitions* if and only if: For all (possibly open) sentences B in the language of

Theory, *Previous-definitions*, and (especially) *Symbol*, there is a (possibly open) sentence B' such that

1. B' is in the language of *Theory* and *Previous-definitions*,
2. *Symbol* does not occur in B', and
3. B and B' are EQUIVALENT relative to *Theory*, *Previous-definitions*, and *Definition*.

Informally: For every sentence *with* the symbol there is an equivalent sentence *without* the symbol — so *all* its meaning is given.

The standard account unpacks the critical notion of EQUIVALENCE in either of two interchangeable ways: proof-theoretically or semantically.

- *Proof-theoretically*: One must be able to prove the "material equivalence" of B and B' from the theorems, previous definitions, and the definition (or their "formal equivalence" if either is open). That is, $B \equiv B'$ (or its universal closure) must be derivable (in a standard calculus) from *Theory*, *Previous-definitions*, and *Definition*.
- *Semantically*: B and B' must be semantically equivalent in the context of the theorems, previous definitions, and the definition; that is, B and B' must have the same truth value in every interpretation of the nonlogical constants (and free variables if present) that renders all of *Theory*, *Previous-definitions*, and *Definition* true.

Equivalence is at the heart of the eliminability. Because equivalence is well understood for truth-functions-cum-quantifiers languages, we can be sure that for these languages the standard account makes sense. If, however, a language does not have a well-understood concept of equivalence, then we cannot be sure what to make of eliminability.

It is plain that the concept of equivalence used in the standard account is an idealization of "inferential equivalence." What is confusing is that this strong notion of equivalence is required even though the sentences that form the premisses and conclusions of the inferences are all assumed to be made from truth functions and quantifiers. In particular, the account would go to pieces with respect to the inferential

use of sentences if one were to substitute "material equivalence" in clause 3 above. This is not surprising since the gap between material equivalence and inferential equivalence is a pretty severe chasm. Here are two witnesses to the cleft. (1) Clause 3 relativizes equivalence to *Theory*, *Previous-definitions*, and *Definition*; but it makes no sense to relativize material equivalence. (2) Suppose, contrary to sense, that clause 3 were stated with material equivalence. To make true the resultant "for every B . . . there is a B' such that . . . B is materially equivalent to B'", we would need only two values of B': a true one and a false one. That would be too easy.

3.4. *Contexts and Meaning*

The discussion of definitions probably goes better if we elaborate the confusing matter of "contexts."

In the first place, there are *explicit contexts* and *implicit contexts*. In the example at hand, all the explicit contexts are provided by the usual array of extensional predicates, extensional operators, extensional connectives (truth functions), and extensional quantifiers leading up to a declarative context. The implicit context is that of inference (or some close cousin). There is in the standard language no "explicitation," in Brandom's phrase, of the inferential context. We have to pay attention to it, and so do the language users, but they have no device by which explicitly to speak of inference. They just do it (so to speak).

In the second place, we can thereby see that there are two contrasts, two different dimensions along which 'context' could be enriched.

With regard to *explicit* contexts, one might have added their modal features or "metalinguistic" features so as to give the users new explicit contexts useful in speaking of inferential connections. This addition might be precisely to explicitate some contexts so far left implicit. The standard account does not deal with these. When such materials are added, we expect an account of meaning that is no longer extensional. We expect new predicates, operators, connectives, and quantifiers that are *intensional* instead of extensional. The general outlines of eliminability and conservativeness remain, but the rigorous details must be elaborated.

With regard to *implicit* contexts, the inferential context imagined in

the standard account is thin stuff. For example, it contains only distant approximations to any of *evidence*, *interest*, *relevance*, *explanation*, or *practical decision*. These are *cognitive* and *practical* concepts, essentially relativized to specific persons with specific mental equipment on specific occasions. The standard account does not provide a notion of equivalence that can sustain replacement in such contexts. Such an equivalence would have to work for *intentional* as well as intensional contexts. It would need to provide us with *sense* equivalence.

Because both the inferential or intensional contexts and the cognitive or intentional contexts are often implicit rather than explicit, it is easy to forget this. Even Tarski, the most methodologically sophisticated definer of all time, once made a mistake about context. The mistake is easy to verify because Tarski 1935 stated his methodology so carefully: For a proposed definition of "truth" to be *adequate*, it is both sufficient and necessary that it be "formally correct" and "materially adequate" in the well known senses that Tarski gave to these phrases (p. 188, including note 1). Evidently Tarski's idea of formal correctness is just what I am calling the standard account. It is equally evident that Tarski's condition of material adequacy suffices to fix the inferential role of "truth," and not just its actual extension, as some folks sometimes say. This is made explicit by his statement in Tarski 1944 that all adequate definitions would have to be "necessarily equivalent" (p. 354). It is anyhow made implicitly clear by his use of the concept of "consequence" in his statement of material adequacy.

The mistake occurs when Tarski suggests an application for his definition. Under the heading of "applicability of semantics to the methodology of the empirical sciences," Tarski first considers the reasonableness of the following.

> An acceptable theory cannot contain (or imply) any false sentences.

His verdict is negative, partly because "we do not know, and are very unlikely to find, any criteria of truth which enable us to show that no sentence of an empirical theory is false" (p. 367).

This application by Tarski is admissible. What makes it so is that the truth-ascription occurs in the above sentence in an extensional context. This is the very sort of context for which the techniques on which

Tarski relies are apt. "Acceptability" is of course a cognitive idea, but the truth-ascription is not within the scope.

Tarski goes on, however, to propose that the following is "an important postulate which can be reasonably imposed on acceptable empirical theory and which involves the notion of truth":

> As soon as we succeed in showing that an empirical theory contains (or implies) false sentences, it cannot be any longer considered acceptable (*ibid.*).

Here Tarski has made a mistake. The difficulty is that in this postulate the truth-ascription does occur within the scope of a (not just nonextensional but) intentional connective. To show this, let me do as best as I can to put the Tarski postulate into middle English.

> For each time t, for each empirical theory T, if (there is a sentence x such that at t we succeed in showing that $((x \in T)$ & $\sim (x$ is true$)))$ then T is not acceptable by us at t.

The mistake is not that the postulate is false; it seems like quite a good thing to say. Nor is it that Tarski's semantic definition of truth cannot help elucidate the postulate, though I certainly think the prosentential theory of truth does a better job of that. The mistake is about definitions. Tarski's mistake is that his definitional methodology fails to license an elimination of "x is true" within the scope of an "at t we succeed in showing that ___." A moment's thought suffices to conclude that inferential equivalence between B and B', which is what Tarski's methods guarantee, could not possibly suffice for replacing B with B' within the scope of this nonextensional and nonintensional cognitive connective. The range of applications for his definition that are secured by his methods, though great, are less than he envisaged.

So the standard theory does not provide for elimination in intentional (as opposed to intensional) contexts. It cannot since its concept of equivalence is itself intensional equivalence (or inferential equivalence) rather than intentional equivalence (or cognitive equivalence, or sense equivalence). The road forks. (1) Some may argue that we want an enrichment of the standard account of definition that would permit elimination in cognitive contexts. (2) Some may argue that we keep to the intensionality-without-intentionality of the standard theory, thereby

giving up a call for elimination in intentional contexts. For a powerful discussion of the issues with particular reference to Frege, see Horty 1993. Horty gives a compelling argument for (2), which is now my view.

These remarks are relevant to the aim of "extensional adequacy" sometimes proposed for a definition (but not by Tarski). Having *attained* such an aim, one has no measure of the inferential use of sentences containing the newly defined *Symbol.* Extension alone does not suffice to specify a role in inference. Given only extensional adequacy, it therefore makes doubtful sense to infer either from or to any sentence with the newly defined *Symbol.* Furthermore, since the practice of assertion is unsnippably tied to inference, we should worry equally about assertion on the basis of mere extensional adequacy.

One might try to base a counter to this line on an idea published in the same year by Bressan 1972 and Kripke 1972, and much earlier in a somewhat different form by Marcus 1961. One may take the newly defined symbol to be "rigid," to adapt Kripke's phrase, or to fall under an "absolute concept," to adapt Bressan's, or to be a "tag," following the idea of Marcus. (These ideas are not the same; absoluteness is deepest of these, and the closely related notion of "substance concept" in Gupta 1980 is equally deep and more manageable. But any will do for the immediate purpose.) Taking this line makes an *indexical* out of the newly introduced symbol, something like "the extension of the definiens as things *actually* stand *here* and *now*." There is nothing even mildly wrong with such a definition; but it cannot be treated as a continuous part of a language, such as that we were envisaging, that is "standard" in being free of indexicals. The only mistake would be to pretend otherwise. In the *standard* language, there is no room for such a counter. Even in the richer language one would find the counter to fail, for careful analysis reveals that you cannot avoid the need for an inferential role, and therefore you cannot avoid awarding an intension. All that happens when a "rigid" term is introduced is that its intension is definitionally determined by its indexically determined extension. My point, however, is not to make theory. It is only this: If we are to think seriously about such definitions, it should be against the background of a careful and rigorous account of them.

3.5. *Criterion of Conservativeness (Rigorous Account)*

9. DEFINITION. (Conservativeness) Let *Definition*, *Symbol*, *Theory*, and *Previous-definitions* be as above. Then *Definition*, qua definition of *Symbol*, satisfies the *criterion of conservativeness* relative to *Theory* and *Previous-definitions* if and only if: For all sentences *B* in the language of *Theory* and *Previous-definitions* (but *not* containing *Symbol*), if *Definition*, *Theory*, and *Previous-definitions* together IMPLY *B*, then *Theory* and *Previous-definitions* (already, without the necessity of *Definition*) together IMPLY *B*.

   Using the language of consequence as a substitute for that of implication, as I usually do below, conservativeness comes to this: If a *Symbol*-free *B* is a CONSEQUENCE of *Theory* and *Previous-definitions* with the help of *Definition*, then it is so without its help.

Informally: Every sentence *without* the symbol is a consequence *without* the definition if it is a consequence at all — so the definition does *nothing but* give the meaning.

The notion of CONSEQUENCE used above can be unpacked in either of two (interchangeable) ways: proof-theoretically or semantically.

- *Proof-theoretically*, consequence means derivability (in a standard calculus).
- *Semantically*, for *B* to be a consequence of some sentences is for *B* to be true in every interpretation making those sentences true.

Just as you cannot substitute material equivalence for equivalence in the preceding discussion of eliminability, so here you cannot substitute material implication for implication or consequence. These ideas make no sense without an inferential context. For example, suppose *Theory* and *Previous-definitions* are empty. Then if we were contrasensically to read "IMPLY" materially, conservativeness would say, I suppose, that for each *Symbol*-free *B*, (*Definition* $\supset$ *B*) $\supset$ *B*, i.e. (*Definition* $\vee$ *B*). And since some *Symbol*-free *B*'s will surely be false, this comes to just: *Definition*. Say that again?

3.6. *Examples*

10. EXAMPLE. Take *Definition* = "for all x, if $x \neq 0$, then for all y, Inverse(x) = y iff $x \times y = 1$."

Thus *Definition* takes Inverse(x) as the standard multiplicative inverse of x, usually written "$1/x$". And Inverse(x) "is defined" only when x is not zero so that the problem of division by zero is avoided. Such a definition is often called a "conditional definition," the *condition* being that $x \neq 0$.

Evidently *Definition* does not satisfy eliminability, since you cannot use it to eliminate all uses of Inverse(0). It does, however, satisfy a "partial eliminability," since *Definition* does permit elimination of Inverse(t) for any term t such that *Theory* and *Previous-definitions* imply that ($t \neq 0$). More generally, *Definition* permits elimination of Inverse(x) in any sentence $\forall x B(x)$ {or $\exists x B(x)$} in which *Theory* and *Previous-definitions* imply that $\forall x B(x)$ is equivalent to $\forall x((x \neq 0) \supset B(x))$ {or that $\exists x B(x)$ is equivalent to $\exists x((x \neq 0) \ \& \ B(x))$}. Since these are known *in advance* as the only contexts of Inverse(t) or Inverse(x) that we care about, such partial eliminability is conceptually satisfying.

We know that Frege thought otherwise; he believed in total eliminability. My own view is that Frege's belief (or practice) arose from a certain slightly misdirected compulsiveness — the same compulsiveness, however, that gave the world its most glorious standard of uncompromising rigor. Many of us, perhaps almost equally compulsive, avoid conditional definitions for what we announce as mere technical convenience. Our alternate device is to use "don't-care clauses," e.g. defining Inverse(0) as equal to 0 (or perhaps as equal to 14), and then remarking that this bit of the definition is a don't-care. If one studies our inferential practices, however, one sees that we never *use* the don't-care clauses, and that we counsel others to avoid relying on them for conceptual interest. So practically speaking, conditional definitions and don't-care clauses come to the same thing.

11. EXAMPLE. (Solubility) Take *Definition* = "for all x, if x is put in water, then x is soluble if and only if x dissolves" as a definition of "soluble."

This is what Carnap called a "reduction sentence." He was surely thinking that *Definition* gave us a "partial" definition of "soluble." Suppose the theory *Theory* contains "Sam is put in water." Then clearly *Definition* permits elimination of "soluble" in the context B = "Sam is soluble," for relative to the theory *Theory*, B' = "Sam dissolves" is equivalent to B. But we couldn't eliminate the defined symbol from "Mary is soluble" unless her presence in water was a consequence of our theory.

The superficial form of a reduction sentence is the same as that of a conditional definition. I think this resemblance misleads many who rely on reduction sentences. But reduction sentences and conditional definitions are very different. In the case of conditional definitions, we do not *want* an account of the newly introduced symbol as applied to arguments not satisfying the condition. In the case of the reduction sentence for solubility, however, the samples whose presence in water we cannot deduce from the theory are *precisely* those samples whose solubility we care about. For example, it would be ridiculous to provide a don't-care clause for the condition, adding the following to Example 11: "and if x is not put in water than x is a rotten apple." That this would be foolish is obvious from examination of the inferential contexts in which one finds the concept of solubility.

I infer that Carnap's terminology represents a serious blunder. Reduction sentences do not give partial "reductions"; unless, on analogy, one wishes to take "counting to three" as "partially counting to infinity."

Still considering *Definition* as in Example 11 above, suppose the following oddity: "t is put in water" is a consequence of *Theory*, for absolutely every closed term t. This would still *not* ensure that *Definition* above satisfies eliminability; for consider some context B = "for every x, x is soluble if and only if x . . . ," where the defined term is used with a variable instead of a constant.

Turning now to the other criterion, the reduction sentence *Definition* of Example 11 is doubtless conservative. Suppose, however, that we enrich *Definition* by adding a conjunct.

12. EXAMPLE. Let *Definition'* be taken as a definition of solubility, where *Definition'* = (*Definition* & $\forall x \forall y$(if x and

y are of the same natural kind, then x is soluble if and only if y is soluble)).

Adopting *Definition'* would doubtless enlarge the cases in which we could eliminate "soluble." But notice that we can now conclude that each thing of the same natural kind as Sam, if put in water, dissolves if and only if Sam does. Unless our *Theory* already committed us to this view, our definition would have smuggled in new assertional content and would not be conservative. A healthy respect for conservation regards such sneaking as reprehensible. Thus again, reduction sentences are not much like definitions after all. They do not reduce, or not enough to count.

What, then, do reduction sentences do? It is perfectly all right to suggest that they describe or establish "meaning relations" between the mystery vocabulary to which "soluble" belongs and the less mysterious language of water and the like. The only thing that is wrong is to think of them as "reducing" the mysterious to the less mysterious, even partially. Only the terminology is wrong.

Here is an example that early Carnap considered on his way to reduction sentences, though I add a moral of which only the later Carnap would have approved.

13. EXAMPLE. (More solubility) Suppose we let *Definition* = "for all x, x is soluble if and only if, if x is put in water, then x dissolves."

This definition of solubility is bound to satisfy both criteria, regardless of theorems of previous definitions. Given our understanding that we are reading English as if it were a standard first order language, it does, however, have the strange (but still conservative) consequence that whatever is not put in water is soluble, so that as *stipulative* it is not a happy choice and as *lexical* it is doubtless false to our usage. One can see here a motive for an alternative theory of "if". The first and best job of following out this insight is to be found in Bressan 1972. In rigorously defining concepts of serious physics, such as the concept of "mass" according to the intuitive ideas of Mach and Painleveé, Bressan uses *modal* concepts precisely because of the failure of extensional concepts to be apt. Carnap, in correspondence with Bressan, strongly encouraged this work.

4. RULES OF DEFINITION

Here, briefly, are the standard rules for producing definitions guaranteed to satisfy the criteria of eliminability and conservativeness. They are easy. And by Beth's Definability Theorem, they are complete for the first order language truth functions with quantifiers. Why is it, then, that so much philosophy otherwise faithful to first-orderism is carried out contrary to the policies they enjoin? Answer: Logic books, excepting Suppes 1957, do not give these matters proper discussion. The consequence is that students of philosophy, even those who are thoroughly taught to manipulate quantifiers, are not taught the difference between acceptable and unacceptable definitions. Since philosophers spend vastly more time proposing and using definitions than they do manipulating quantifiers, this is sad.

As promised, the rules are presented as definitions. (Surely it won't occur to anyone to think in this regard of circularity.)

14. DEFINITION. (Standard rule for defining by an equivalence) *Definition is a standard definition of a predicate symbol R relative to Theory and Previous-definitions* iff for some n (the n-arity of R), and some variables $v_1, \ldots, v_n$, and some sentence A (the "definiens"),

    1. *Theory* and *Previous-definitions* are sets of sentences (theory and previous definitions, respectively), R is an n-ary relation symbol (to be defined), and *Definition* is a sentence (the candidate definition).
    2. *Definition* is an n-times universally quantified biconditional $\forall v_1 \ldots \forall v_n(Rv_1 \ldots v_n \equiv A)$.
    3. The variables $v_1, \ldots, v_n$ are distinct.
    4. The definiens A has no "dangling" or "floating" variables, that is, no *free* variables other than $v_1, \ldots, v_n$.
    5. Each nonlogical symbol in A is drawn from among those of *Theory* and *Previous-definitions*.
    6. R is foreign to *Theory* and to *Previous-definitions*.

It follows that in particular, R does not occur in A.

15. DEFINITION. (Standard rule for defining an operator by an

equivalence) *Definition is a standard definition of an operator symbol O relative to Theory and Previous-definitions* iff for some n (the n-arity of O), and some variables $v_1, \ldots, v_n$, w, and some sentence A (the "definiens"),

1. *Theory* and *Previous-definitions* are sets of sentences (theory and previous definitions, respectively), O is an n-ary operator symbol (to be defined), and *Definition* is a sentence (the candidate definition).
2. *Definition* is an $(n + 1)$-times universally quantified biconditional $\forall v_1 \ldots \forall v_n \forall w((Ov_1 \ldots v_n = w) \equiv A)$.
3. The variables $v_1, \ldots, v_n$ and w are distinct.
4. The definiens A has no free variable other than $v_1, \ldots, v_n$ and w.
5. Each nonlogical symbol in A is drawn from among those of *Theory* and *Previous-definitions.*
6. O is foreign to *Theory* and to *Previous-definitions*
7. The sentence $\forall v_1 \ldots \forall v_n \exists y \forall w[(y = w) \equiv A]$ is a consequence of *Theory* together with *Previous-definitions.* That is, it is a consequence of the theory together with previous definitions that, for each n-tuple of arguments $v_1, \ldots, v_n$, there is in fact exactly one value w such that A; so A is "functional."

So O does not occur in A.

5. ADJUSTING THE RULES

Even within the purview of the standard account, there are other forms of definition with an equal claim to propriety. In the context of the axiom of extensionality in set theory, for instance, one may properly define set-to-set operators by way of membership conditions. In the context of higher order axioms establishing an inductively generated structure (such as Peano's axioms), one may properly define operators by rehearsing the mode of inductive generation. Sometimes, in some contexts, some of these are called "implicit definitions." In these cases, "properly" means "in such a way as to satisfy the criteria of eliminability and conservativeness." So much is a built-in generality of the standard account.

Suppose, however, the inductive material is added instead in a first order way, including a license for new inductive definitions. People who wish to avoid reliance on higher order logic sometimes proceed in this fashion. Expressions so defined are likely to be inferentially enriching and thereby violate the criterion of eliminability. This inductive practice is so well understood, however, that it appears unreasonable to call it into question — unless perhaps the question is foundational. Otherwise it may be better to weaken the criteria and bend the rules.

If the language at issue is changed or extended, with the implication that eliminability is now required for additional contexts, then the rules need adjusting in another sense; that is, following the rules may not suffice for satisfying the criteria. For instance, a universally quantified biconditional as in Definition 14 will evidently not permit elimination of a defined predicate from a modal context. Should, then, *Definition* have the form $\Box\forall x(Rx \equiv A)$, or should it instead have the form $\forall x\Box(Rx \equiv A)$? Or will both or neither of these do? The fundamental point is this: We do not have to resort (much) to intuition. It is a matter of what is required to satisfy the criteria; that's all.

Relevance logic offers an excellent example of the same phenomenon. It is an interesting exercise to work out how to adjust the rules for defining new operators in Meyer's relevant arithmetic (see Anderson, Belnap and Dunn 1992). With one eye on the two criteria, you can see with the other precisely why and in what sense relevant arithmetic does and should prohibit that a division operator be defined.

A language with explicit tolerance for vagueness unquestionably needs its own matching story about definitions. Some of Tappenden's forthcoming work on "pre-analytic" statements seems to me to count as in the required direction.

6. RELAXING THE CRITERIA (OR CHANGING THE RULES)

The standard criteria and rules must be counted as logic of the greatest interest. They are, I should say, always to be respected as devices to keep us from addle. On the other hand, one can easily find cases in which good thinking bids us moderate their force. Here is an illustration concerning noncircularity.

Gupta (1988/1989) established the following.

1. Some concepts are essentially circular. (Normal results of inductive definitions, e.g. multiplication, are *not* examples of circular concepts.)
2. The standard account of definitions says nothing useful about circular concepts (I suppose it denies their existence).
3. One obtains a powerful theory of circular concepts by reworking the theory of definitions to admit circular definitions.
4. Truth (in e.g. English) is a circular concept.
5. The ideas of the reworked definitional theory, when applied to truth, make fall into place both the ordinary and the extraordinary (pathological, paradoxical) phenomena to which philosophers have called attention.

These notions are extended and defended in *The Revision Theory of Truth*, Gupta and Belnap 1992, which I will reference as RTT. It is not possible to summarize even the main ideas here; I only make a few remarks relating the RTT work to the standard account of definitions.

RTT fully *accepts conservativeness*. The norm remains: If it's a definition, then it should not provide camouflage for a substantive assertion. I shan't say anything more on this topic.

RTT *abandons eliminability, and in particular noncircularity*. The indispensable idea is that a definitional scheme cannot illuminate a circular concept except with the help of a circular definition, and that circularity intrinsically prevents eliminability. As the standard account is satisfied with only partial eliminability for conditional definitions, so the RTT account is satisfied with only partial eliminability for circular definitions. For example, the RTT definition of truth permits its elimination from " 'Euthyphro is pious' is true," but not from "This very sentence is true." There is, however, a deep difference between how noncircularity is relaxed in the presence of standard conditional definitions and in the presence of RTT circular definitions. In the case of the former, there is an inferential account of when eliminability is to be expected, and when not; the very form of a standard conditional definition supplies material for such an account, as explained above. In contrast, on the RTT theory of circular definitions, eliminability may depend on contingent facts. There is no grammatical or even inferential

test to distinguish eliminable, innocuous circularity from ineliminable, vicious circularity. Furthermore, RTT also abandons the other support of eliminability, namely, no inferential enrichment. In the case of truth, RTT only requires that a definition fix for it an appropriate intension or inferential role. It does not require that there already be present in the language something else that occupies that role. RTT thinks of truth as an enrichment of the language in that sense, but as "definable" in the sense that its intension is fixed. Adding truth is therefore just like adding The Absurd, as in Example 5, to a purely positive language. In both cases the inferential role is uniquely determined, but unoccupied prior to the definition.

RTT *changes the grammar* of definitions. Definitions are not given explicitly as sentences, or even as theories, of the language. Instead, "implicit" definitional ideas are introduced. In one scheme, rather than requiring *Definition* to be a sentence, as in Definition 14, a definition straightway relates a definiendum to a definiens. The system of definitions is thought of as an implicit part of the language in just the way that inferential connections are implicit. We, standing outside the language, may clearly describe a part of such a system by writing e.g. "$Rx =_{\mathrm{Df}} A$", just as we may write "A is a consequence of B." Other rules remain in force, except, crucially, (6) of Definition 14.

16. EXAMPLE. RTT ignores an explicit Tarski biconditional such as " 'Euthyphro is pious' is true ≡ Euthyphro is pious," but it takes very seriously that there be a nonsymmetric definition relation between " 'Euthyphro is pious' is true" and "Euthyphro is pious."

    In an exactly parallel fashion, where it is assumed that *Liar* = "*Liar* is not true", RTT ignores the self-contradictory sentence "*Liar* is true ≡ *Liar* is not true", but it takes with great seriousness the nonsymmetric definitional relation between "*Liar* is true" and "*Liar* is not true".

RTT *provides a semantic account* of circular definitions. The Kernel idea is given by the word "revision." You start with an extension for all but the various definienda, and you choose a hypothesis as to what their extensions might be. You now have a hypothetical extension for

everything in the language, and therefore for every definiens. You use the definitions to *revise* your hypothesis, arriving at a new hypothetical extension for each definiendum. And so forth — a phrase here indicating progress into the transfinite. The hope is that you can eventually arrive at categorical results independent of your starting hypothesis.

An especially fascinating part of this story is that RTT furnishes something worthwhile to say about circular concepts such as truth even in those cases in which eliminability fails. It does so by attending to their *patterns of revision* under different hypotheses. In this way forms of pathologicality can be distinguished one from the other in a conceptually satisfying way. You can *see* the difference between "This very sentence is true" and "This very sentence is false" by watching the unfolding of their altogether different patterns of revision.

RTT *describes proof-theoretical techniques* corresponding as closely as might be hoped to the semantics of revision, thus completing the trio grammar/semantics/proof theory. Of particular interest is that the proof theory reflects a fundamental distinction between introducing and eliminating a term by means of a definition.

7. CODA

Two items remain. The first is to express regret that these remarks have been too coarse to capture the fine structure of even the standard account of definitions. The second is to make explicit that the general lesson has the usual banal form. On the one hand, the standard criteria and rules are marvelous guides. They have point. They are not to be taken lightly. Every philosopher and indeed I should say everyone engaged in conceptual work ought to know and understand them. He or she who unknowingly breaks with the standard criteria or the rules risks of folly. On the other hand, there is no single aspect of these criteria or rules that we should consider unalterable.

NOTE

[1] Thank to A. Gupta and J. Tappenden for many-sided help.

REFERENCES

Anderson, A. R., Belnap, N. D. and Dunn, J. M. (1992) *Entailment: The Logic of Relevance and Necessity*, Vol. 2, Princeton, Princeton University Press.

Bochenski, J. (1956) *Formale Logik*, Freiburg, K. Alber. Translated and edited by I. Thomas, *A History of Formal Logic*, second edition, New York, Chelsea Publishing Company, 1970.

Bressan, A. (1972) *A General Interpreted Modal Calculus*, New Haven and London, Yale University Press.

Carnap, P. (1937) *The Logical Syntax of Language*, International library of psychology, philosophy and scientific method, New York, Harcourt, Brace. Translation of *Logische Syntax der Sprache* by A. S. Smeaton (Countess von Zeppelin).

Church, A. (1956) *Introduction to Mathematical Logic*, Princeton, NJ, Princeton University Press.

Couturat, L. (1905) "Les définitions mathématiques," *L'Enseignement mathématique* 7, 27—40.

Curry, H. (1963) *Foundation of Mathematical Logic*, New York, McGraw-Hill, Inc.

Frege, G. (1964) *The Basic Laws of Arithmetic; Exposition of the System*, Berkeley, University of California Press. Translated and edited, with an introduction, by M. Furth. The translation is of "the introductory portions of the first volume and an epilogue appended to the second."

Gupta, A. (1980) *The Logic of Common Nouns: An Investigation in Quantified Modal Logic*, New Haven, Yale University Press.

Gupta, A. (1988/1989) "Remarks on Definitions and the Concept of Truth," *Proceedings of the Aristotelian Society* 89, 227—246.

Gupta, A. and Belnap, N. (1992) *The Revision Theory of Truth*, Cambridge, MA, The MIT Press. In press.

Horty, J. (1993) "Frege on the Psychological Significance of Definitions," *Philosophical Studies* 72. 223–263 (this issue).

Kneale, W. and Kneale, M (1962) *The Development of Logic*, Oxford, The Clarendon Press. Reprinted from corrected sheets of the first edition, 1966.

Kneebone, G. T. (1963) *Mathematical Logic and the Foundations of Mathematics*, Princeton, NJ, New York, Toronto, London, D. van Nostrand Publishing Company.

Kripke, S. (1972) "Naming and Necessity." In: D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, pp. 253—355. Dordrecht, D. Reidel Publishing Company.

Leśniewski, S. (1931), "Über Definitionen in der sogenannten Theorie der Deduktion," *Comptes rendus des séances de la Société des Science et des Lettres de Varsovie*, Classee III, 24, 289—309.

Leśniewski, S. (1981) *Collected Works*. S. J. Surma, J. T. J. Srzednicki, and D. I. Barnett (eds.), Synthese library, vol. 118, Boston, D. Reidel Publishing Company.

Luschei, E. C. (1962) *The Logical Systems of Leśniewski*, Amsterdam, North-Holland Publishing Company.

Marciszewski, W. (1981) "Definition." In: Marciszewski, W. (ed.), *Dictionary of Logic*, pp. 89—96. The Hague, Martinus Nijhoff.

Marcus, R. B. (1961) "Modalities and Intensional Languages," *Synthese* 13, 303—322.

Robinson, R. (1950) *Definition*, Oxford, The Clarendon Press.

Runes, D. D. (1962) *Dictionary of Philosophy*, Totowa, NJ, Littlefield, Adams.

Suppes, P. (1957) *Introduction to Logic*, Princeton NJ, New York, Toronto, London, D. van Nostrand Publishing Company.

Tarski, A. (1935) "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophica* 1, 261—405. German translation of a book in Polish, 1933. English

translation by J. H. Woodger, "The Concept of Truth in Formalized Language," in *Logic, Semantics, Metamathetics*, papers from 1923 to 1938 by A. Tarski, Oxford, Oxford University Press, 1956, pp. 152—278.

Tarski, A. (1941) *Introduction to Logic and to the Methodology of Deductive Sciences.* Enlarged and revised edition, New York, Oxford University Press. Translated by Olaf Helmer from a work that appeared in Polish in 1936 and the in German translation in 1937.

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

Tarski, A. (1956) *Logic, Semantics, Metamathematics*, London and Oxford, The Clarendon Press.

Department of Philosophy
University of Pittsburgh
Pittsburgh, PA 15260
USA