J. D. Fodor, J. A. Fodor, M. F. Garrett

The Psychological Unreality of Semantic Representations

It is one of the consolations of philosophy that the benefit of showing how to dispense with a concept does not hinge on dispensing with it. W. V. Quine on definition without elimination.

1. Introduction

As everybody knows, the treatment of semantics within generative grammars has been controversial even by the standards prevalent in linguistics. Neither the properties of semantic representations nor the way they relate to syntactic structures is generally agreed upon, though these are surely among the basic issues that any adequate semantic theory must decide.

However, some consensus is to be found: much of the recent literature assumes at least implicitly that the grammar must contain a semantic level that satisfies the following three conditions.

A. Semantic representations are psychologically real in the sense that, given appropriate idealizations, understanding a sentence requires the recovery of its semantic representation.

B. A variety of properties and relations that are pretheoretically identified as semantic are formally definable over representations at the semantic level. These have traditionally included synonymy, anomaly, and ambiguity, but in recent years discussion has come to center on entailment (Katz 1972, Lakoff 1970). In particular, something like the following is supposed to be a constraint on the semantic level: if, according to the intuitions of speaker/hearers, sentence $S_1$ of language $L$ entails sentence $S_2$ of $L$, then the argument from the semantic representation of $S_1$ to the semantic representation of $S_2$ must be formally valid.

C. Not every lexical item of the natural language corresponds to a syntactically simple expression at the semantic level. That is, there are definable lexical items of $L$, and these are to be represented by defining phrases of the language in which semantic representations are couched. A standard example is the English word bachelor, which is supposed to be defined at the semantic level by some such phrase as ‘unmarried, adult,
male human’. An intended consequence of this is that sentences that are related in the way that *John is a bachelor* is related to *John is an unmarried man* should have identical representations at the semantic level.

The major contention of this article is that there are persuasive empirical grounds for believing that no level meeting conditions B and C will also meet condition A. This is to say that the semantic level, as linguists have generally conceived of it, does not exist. Before considering the evidence for this claim, we need to review a number of points about the status of conditions A–C and their relations to other assumptions about the character of generative grammars.

1.1. Condition A

The sort of psychological reality claim that A expresses is presumably not specific to the semantic level. In fact, those who hold that the semantic level ought to satisfy A usually do so because they endorse the psychological reality of structural descriptions at all the levels that the grammar postulates. This condition is presumably the weakest that a grammar can satisfy if it is to contribute to a psychological model of the speaker/hearer at all. Thus, for example, claims for the psychological reality of grammatical operations entail claims for the psychological reality of structural descriptions, but not vice versa. (For further discussion, see Fodor, Bever, and Garrett 1974).

It is important to emphasize, however, that the particular psychological status attributed to semantic representations by A does not follow from the mere claim that semantic representations are psychologically real. Condition A assigns semantic representations a role in the comprehension of sentences, but it is possible to imagine that they play no role in that process but are engaged instead by some other psychological states or computations. As far as we can see, however, memory for sentences would be the only other plausible candidate, and the available evidence seems to show decisively that what is recalled in remembering a sentence corresponds to none of its linguistically motivated representations. (See Bartlett 1932, Bransford, Barclay, and Franks 1972, Johnson-Laird and Stevenson 1970 and the discussion in Fodor, Bever, and Garrett 1974).

Moreover, the view that semantic representations are implicated in the sentence comprehension process is independently plausible since it provides for an extremely natural account of communication exchanges between speakers and hearers. On this account, the formal objects that are encoded and recovered in speech exchanges are semantic representations. It seems that any psychological model of such exchanges must recognize some formal object which captures the notion of the message standardly communicated by uttering a sentence. The view we are considering here—which, in fact, we endorse—requires that this object be among the structural descriptions that the grammar assigns to the sentence. If grammars do contain a level meeting conditions B and C, then representations at that level are the obvious candidates for
the job of specifying messages. In other words, it would be distinctly strange (though not logically incoherent) to recognize a linguistic level meeting conditions B and C that does not meet condition A.

1.2. Conditions B and C

These conditions, unlike A, are intended to be proprietary to the semantic level. Moreover, they are intimately interrelated, since C provides one of the means for satisfying B. For example, if, at the semantic level, definable words are replaced by their definitions, then the synonymy relation can perhaps be reconstructed in terms of identity of semantic representation (Katz 1972). Condition C is also presumed to permit inferences turning on “content” words of the language to be captured in essentially the same way as inferences that turn on “logical” words like and, or, if, some, all, etc. Thus, the argument from John is a bachelor to John is unmarried is reconstructed at the semantic level as an argument from 'John is unmarried and adult and male and human' to 'John is unmarried',¹ which is itself an instance of the valid inference scheme \( a \text{ is } F \& G \rightarrow a \text{ is } F \).²

This brings us to the third point about A–C, namely that they are neutral in the controversy between generative and interpretive semantics since they are presupposed by both. Both theories appeal to eliminative definitions to capture the validity of inferences that turn on content words. Viewed from this perspective, the major dispute between these schools is over how, not whether, defining expressions are assigned to definable lexical items in the natural language.

This is not to claim that the issues between generative and interpretive theories are exhausted by their respective accounts of the lexicon, or even that either kind of theory is inherently committed to eliminative definitions. Chomsky’s recent accounts of the “extended standard theory”, for example, propose what is clearly an interpretive model of the relation between syntactic and semantic representations, but they in no way rely upon the decomposition of lexical items. The remarks that follow are intended to apply only to those theories that do employ eliminative definitions.

Generative and interpretive theories agree that a dictionary is one of the essential components of a semantic theory. In the sense they have in mind, a dictionary consists of a finite set of entries, where each entry is an object language lexical formative together with its defining expression. According to interpretive theories, the input to the semantic component consists of syntactic structures drawn from one or more levels of the grammar. The process of constructing semantic representations for object

¹ We shall put semantic representations in single quotation marks. Presumably semantic representations are couched in the vocabulary of some universal metalanguage. Since, however, we have no idea of what that language looks like, we will allow English expressions to stand as surrogates for its formulae. However, nothing in the discussion to follow turns upon our doing so.

² We are not, of course, imputing the details of this example to all—or, indeed, any—of the semantic theories that have been developed in the recent linguistic literature. Even in such simple cases, a variety of different assumptions is possible concerning the nature of the defining expressions and the rules of inference that apply to them. Current theories do, in fact, differ on just such matters.
language sentences exploits the dictionary to replace the formatives in these syntactic structures by their defining expressions. The rules for applying the dictionary are sui generis and in particular do not satisfy constraints on syntactic transformations. In a generative semantic theory, by contrast, the rules that relate object language formatives to their definitions are construed as a special case of substitution transformations, and are thus required to meet the usual conditions on such transformations. Semantic representations thus constitute the domain of transformational operations, and object language formatives are introduced by transformations which, in effect, substitute them for their definitions. In short, generative and interpretive theories differ on where, in the course of derivations, rules of definition operate, and on the formal character of such rules. However, there seems to be a general consensus that grammars contain some rules of eliminative definition.

But despite the central role that has been accorded to definitions in so much recent semantic theorizing, alternative approaches to word meaning have, from time to time, been suggested (Carnap 1956; Bar-Hillel 1967), and linguists have occasionally taken these suggestions seriously as supplements to the classic definitional approach (Fillmore 1971; Lakoff 1970). In what follows we shall argue for abandoning the definitional approach entirely since, so far as we can tell, it is inherently unable to account for a variety of pertinent facts about the way in which people understand sentences.

2. Some Psychological Evidence

In this section, we will review two kinds of considerations that seem to us to militate against the psychological reality of eliminative definitions, and hence against the existence of a level of grammatical representation that simultaneously satisfies A and C. The first of these considerations is as yet only a matter of informal observation; the second turns on some experimental results.

2.1. Intuitive Evidence

Theories that assume that the application of eliminative definitions is a prerequisite to understanding sentences appear to be committed to certain intuitively implausible predictions about sentence complexity: in particular, that, all other things being equal, the relative complexity of a pair of sentences should be a function of the relative complexity of the definitions of the words that the sentences contain. Thus, to pursue our very simple example, the sentence John is a bachelor should be more complex than the sentence John is unmarried, given that the semantic representation of the former properly includes the semantic representation of the latter. And, by a similar argument, Floyd broke the glass should be more complex than Floyd saw the glass if, as has been proposed, causative verb constructions are analyzed, at the semantic level, into structures containing (at least) two clauses, one of which has an explicit verb of
causing and the other an inchoative verb. Again, *Cats chase mice* ought to be more complex than *Cats catch mice* if, as Katz (1972) has suggested, the semantic analysis of *chase* involves its decomposition into some structure that includes explicit reference to an intention to catch.

Which predictions one makes about the relative complexity of particular pairs of sentences depends, of course, on which definitions one assumes for the language in question. But any definition that is postulated should result in asymmetries of the kind just mentioned: the semantic representation of a sentence containing the defined term will possibly require more operations to recover, and will certainly require more memory space to display, than that of a corresponding sentence containing only a part of the definition. Experimental results addressed to this question are hard to come by, so, as things stand, one is forced to rely largely on intuition. However, it is demonstrable (Schwartz, Sparkman, and Deese 1970) that intuitions of relative syntactic complexity are reliably correlated with experimentally derived rankings, so they have a prima facie claim to construct validity in the present case too. And they do not appear to support the asymmetries of semantic complexity that definitional theories predict.

It might be supposed—it might even be the case—that if the predicted asymmetries are not actually observed, that is only because they are too small to be registered by linguistic intuition. However, it should be noticed that in the kinds of cases cited above, the difference between the members of a sentence pair runs to a whole underlying clause. It does not seem unreasonable to suppose that differences of this magnitude ought to have some influence upon intuitions of relative sentence complexity. There are also standard candidates for definitional relations where the difference between surface sentences of intuitively comparable complexity would amount to several clauses at the level of semantic representation. Thus, for example, the Russellian eliminative definition of *the* renders *The man I met is bald* as $(\exists x) (I\ met\ x\ and\ x\ is\ a\ man\ and\ x\ is\ bald\ and\ (y)\ (if\ y\ is\ a\ man\ then\ (either\ (I\ didn't\ meet\ y)\ or\ (y \ = \ x))))$. (Roughly: 'I met a man and he is bald and I met no man other than him.') On the other hand, *A man I met is bald* is analyzed as $(\exists x) (I\ met\ x\ and\ x\ is\ a\ man\ and\ x\ is\ bald)$. If, then, we assume that these definitions are both displayed at the semantic level, we must predict a difference of three clauses in the semantic representation of *The man I met is bald* vs. *A man I met is bald*, a difference of six clauses in *A man hit a ball* vs. *The man hit the ball*, and so on. Patently, intuition does not support any such differences in complexity.

One might, quite plausibly, seek to avoid this kind of embarrassment by suggesting that *the* and *a* should both be treated as primitive (that is, unanalyzed) expressions at the semantic level. The Russellian schema would then function not as an eliminative definition in the sense of B and C, but rather as a "meaning postulate" in approximately the sense of Carnap (1956). But if one is prepared to rule out eliminative definitions by such a move in one case, the question immediately arises why one
should not do so in every case, thereby removing definitions from their present central role in the system that assigns semantic representations to sentences. Of course, this would have the consequence that grammatical theory would not recognize a semantic level of the kind characterized by B and C. It is, in fact, our view that the failure of correlation between definitional and psychological complexity is so considerable as to suggest that this is the right move to make. We shall return to this later in the article.3

2.2. Experimental Evidence

The second sort of consideration that militates against the psychological reality of eliminative definitions is really just a special case of the failure of correlation between definitional and intuitive complexity just noted. In this case, however, differences in complexity are predicted not on the basis of the number of elements in semantic representations but on the basis of which elements are present. In particular, we will be concerned with the contribution of negative elements to determining the intuitive complexity of sentences.

Prima facie, there are four different sources for the element ‘negative’ in a semantic representation, i.e. there are four kinds of lexical formative that may contribute a negative element to the semantic representation of sentences that they occur in. To begin with, ‘negative’ may be inherited from an explicitly negative free morpheme; *not* is an obvious example. Second, ‘negative’ may be inherited from a bound morpheme such as *in-*, *un-*, *im-*, etc. We shall call words containing such morphemes morphological negatives. Third, ‘negative’ may be inherited from an implicitly negative morpheme, viz. one that is negative in sense and whose scope exhibits such typical reflexes of negation as *any*, *much*, *give a damn*, etc., but which has no obvious morphological decomposition in terms of explicit negation. *Doubt*, *deny*, *fail*, etc. are paradigm examples of this class, though the border between morphological and implicit negation is hard to draw: *neither* is a morphological negative if its morphological analysis is *not* + *either*, but an implicit negative if it is morphologically simple. For present purposes, however, the issue is not how to distinguish the first three classes from each other, but rather the existence of a fourth source of ‘negative’ in semantic representation that is clearly distinct from any of these others. We have in mind here words that have

3 Since this was written, Professor Edward Martin has called our attention to the work of Kintsch (1974). Kintsch performed a number of experiments that were designed to test the relative psychological complexity of sentences that differed in the definitional complexity of their constituent lexical items. Among the experimental procedures he employed was the “phoneme monitor” task (see Foss 1969), which has generally proved a reliable instrument for measuring differences in the difficulty of understanding sentences. Kintsch’s results uniformly failed to exhibit the asymmetries of complexity that definition-based theories of semantic representation predict. On the basis of these findings, Kintsch argues that meaning postulates are a more plausible mechanism than definitions for reconstructing the contribution of items in the nonlogical vocabulary to the determination of entailment relations between sentences. His view is fundamentally similar to the one that we will develop here, though there are significant differences of detail.
'negative' as an element of their definitions, but which contain no negative morpheme, and which do not constitute syntactic environments for negative polarity items and so are not implicitly negative by that criterion. As it happens, several of the classic candidates for definitional analysis belong to this fourth class. Thus bachelor is supposed to mean 'man who has not ever been married', kill has been claimed to mean 'cause to become not alive', and on at least one standard representation of the meaning of the there is a negated clause in the "uniqueness" condition.

Let us call items of the fourth type pure definitional negatives (PDNs). In doing so, we stress that if 'negative' appears in their linguistic representations at all, it does so only at a level where definitions replace lexical items. The interest of PDNs is that they constitute a test case for the existence of such a level; if it can be demonstrated that their linguistic representations do contain 'negative', that would argue for the psychological reality of definitions, and hence for the existence of a semantic level in the sense of conditions A–C. Conversely, if it can be shown that their linguistic representations do not contain 'negative', that would argue for the unreality of definitions, and hence for the nonexistence of a semantic level of the kind that A–C postulate. We will argue that the linguistic representation of PDNs does not, in fact, contain 'negative'. If this is right, then there are at least some morphemes which are definable in principle but which must be undefined at any level of grammatical representation with claims to psychological reality.

There exists a litmus that can be used for detecting the negativity of a linguistic item: namely, such items typically interact with quantifiers and with each other to produce a noticeable increase in the complexity of a sentence. This is true not only for explicit negatives (as in It's not true that nobody came) but also for implicit negatives (as in It's false that they denied that he was ill). We have recently conducted an experiment which seeks to use this phenomenon to determine whether there is a negative element in the linguistic representation of PDNs. The detailed results of this study will be reported in a later publication, but the general outline can be given here.

Subjects were requested to evaluate the validity of a number of arguments, each of which contained a quantifier or explicit negative element together with either (a second) explicit negative, or a morphological negative, or an implicit negative, or a PDN. The stimulus materials were so arranged that performance on each of the three latter kinds of arguments could be compared with performance on an argument of the first kind. That is, the result allowed us to evaluate explicit negative vs. morphological negative, explicit negative vs. implicit negative, and explicit negative vs. PDN. In each of these comparisons, the paired arguments differed only with respect to the negative formatives they contained. (So, for example, the comparison between explicit negative and PDN might be made with respect to such sentences as: If practically all of the men in the room are not married, then few of the men in the room have wives.) The dependent variable was reaction time to a correct evaluation of the validity of the argument.
The results may be summarized as follows.

1. Both morphological and implicit negatives tended to be easier than their explicit counterparts, though in neither case is this tendency statistically significant on our current data.

2. The difference between morphological negatives and explicit negatives is quite comparable to the difference between implicit negatives and explicit negatives. Nothing in the results distinguishes subjects' performance on morphological and implicit negatives.

3. Arguments containing PDNs, however, were significantly easier than the paired arguments containing explicit negatives. Moreover, and most important, the difference between PDNs and explicit negatives was significantly greater than the difference between explicit negatives and either implicit or morphological negatives. We take this result to suggest strongly that PDNs do not act as though they contain a negative element in their linguistic representation; and therefore, that PDNs are not semantically analyzed at any level of linguistic representation.

3. Overview and Conclusions

We have now considered two lines of evidence suggesting that there is no semantic level meeting conditions A–C; in particular, that no level that meets B and C is psychologically real. Of course, we do not view this evidence as literally establishing that claim. But it seems to us sufficiently persuasive to warrant raising the question of what the grammar and the psycholinguistics might look like if we assume that there is, in fact, no such level.

So far as we can see, there are three broad options.

1. We could hold to B and C while abandoning A. That is, we could give up the claim for the psychological reality of structural descriptions, either in general or for the special case of semantic representations. Clearly there is nothing in principle to stop

4 It might be suggested that since implicit negatives are not significantly different from explicit ones, the former are negative by the complexity tests; and that this would require that some words have definitions even if some others (PDNs) do not. However, this argument is inconclusive. The fact that words like doubt and deny exhibit the characteristic interactions with quantifiers and overt negative elements shows that they must somehow be classed with negative words by the grammar. One way to so classify them would be to decompose them into a negative element plus further content material (so that doubt becomes ‘not believe’, deny becomes ‘state that not’, etc.). An alternative treatment, however, is simply to acknowledge a class of negative primitives that includes both the explicit and the implicit negatives.

In short, the fact that implicit negatives behave like explicit negatives in the relevant respects is compatible with representing the latter as containing negative elements at some level, but it does not literally require that treatment. On the contrary, one might argue that given some words (PDNs) that could be defined but are not, one should opt for the nondefinitional treatment of implicit negatives as well. Otherwise, we might have to tolerate an uninterpreted degree of freedom in the linguistic theory. Clearly, we should prefer a theory that requires all definables to be defined or a theory requiring that none of them be, as compared to a theory that allows definitions just when it happens to feel like doing so. At the very least, a “mixed” theory would need some principled distinction between the words that are defined and the ones that are not. (For an attempted defence of a mixed theory, see Lakoff 1970.)
one from taking this line, since it amounts merely to weakening the empirical conditions by which grammars have usually been constrained. However, anyone who argues this way owes an account of what is at issue between competing claims about the structural descriptions of sentences, and how, even in principle, such claims can be resolved.

On the view that structural descriptions are psychologically real, the answers to such questions are clear in principle (even though hard to establish in practice). Hypotheses about structural descriptions are true when they correspond to the relevant facts, and the relevant facts concern the internal representations that speaker/hearers compute when they produce and understand the sentences of their language. But if it isn’t internal representations that make structural descriptions true, what is it?

It is becoming fashionable to try to answer this question by viewing the grammar merely as a representation of a certain abstract object—the language—and characterizing truth for the grammar as correspondence between the statements that it makes and the properties of that object. The intended analogy is to the purely formal sciences like logic and mathematics, since it is claimed that what makes theories in such sciences true is a correspondence between the statements that they make and the properties of abstract objects such as numbers, propositions, and functions. In particular, a formal science like mathematics makes no claims for the psychological reality of the entities it describes.

The difficulty with this position, however, is that it seems to make the empirical methodology of linguistics virtually unintelligible. Linguists normally take the intuitions of speaker/hearers to be the data to which structural descriptions are required to respond. But this practice would be quite unwarranted unless it were assumed that speaker/hearers do have access to internal representations of sentences and that these provide a reliable source of information about the character of the abstract object (the language) which, on any view, the grammar is ultimately intended to describe.

In short, the primary data of linguistics are psychological data. Purity of method would thus suggest either that we use no psychological information, including intuitions, to constrain the grammar, or that we use all the pertinent psychological information to constrain it. However, the former alternative is not a serious option, since it would, in effect, limit the empirical data for linguistics to regularities in the corpus, which, as everyone now knows, hopelessly underdetermine grammatical descriptions. On the other hand, the latter alternative makes both intuitive and experimental data germane, in principle, to the validation of existence claims for linguistic levels, and this brings us back to the view that structural descriptions are intended to be psychologically real.5

5 Of course, we do not take all linguistic intuitions to be veridical. Nor need all the experimental data which seem prima facie relevant to constraining grammars turn out to be so. In some cases, we may conclude that the observations are not captured directly by the grammar, but rather by the interaction of the speaker/
2. The second possibility is to claim that conditions A–C would define a linguistic level if there were any arguments whose validity turned on the meaning of terms in the nonlogical vocabulary, but that there are in fact no such arguments. This would be to embrace an austere notion of validity such that the only deductively valid arguments are those that turn on the meanings of the logical words. The argument from is a bachelor to is unmarried is then not deductively valid; that bachelors are unmarried is just a contingent truth. This approach preserves B and C since it takes them to be vacuously satisfied, but it does so by sacrificing all the textbook examples of lexical decomposition.

Many linguists find this sort of position so counterintuitive as to be untenable. But it should be noted that, whatever the case may be with bachelor and unmarried, a number of other classic examples of putative truths of definition (e.g. Cats are animals) have recently come under attack on epistemic grounds of the familiar Quineian variety (cf. Putnam, forthcoming). The point of these arguments is that scientific advances might lead us to abandon such statements, and that this shows them to be empirical rather than necessary. And, presumably, if they are not necessary, then they cannot be true by definition.

We cannot engage in this debate here. For present purposes, we will assume that provision does need to be made for arguments whose validity turns on the meaning of content words, and hence for some mechanisms which will do what B and C do in standard semantic theories. But we allow that it may turn out that there are very few such arguments, and hence little work for this mechanism to do.

3. The third option requires some preliminary exposition.

One way of understanding conditions A–C is to see them as incorporating a certain claim about the relation between understanding a sentence and determining what it entails. Thus, if A–C are true, then merely understanding John is a bachelor amounts to recognizing that it entails that John is an unmarried man. This is because understanding a sentence requires recovering its semantic representation, and, by hypothesis, the semantic representation of John is a bachelor is ‘John is an unmarried man’. No further inferential principles are required to capture the entailment.

Contrast this case with the inference from John is a bachelor to the conclusion that hearer's internal representation of his language with other psychological entities, states, and processes. But such conclusions need positive justification; merely appealing to a “performance/competence” distinction does not absolve the linguist from the obligation to account for the data. So, in the present case, if the sorts of evidence we have cited are judged compatible with the existence of a traditional semantic level, one needs an explanation which shows how they are compatible, and one needs some independent evidence for accepting that explanation.

The situation is complicated by the consideration that the Quinean argument is directed, in the first instance, not against lexical decomposition per se but rather against the necessity of such statements as Cats are animals. It is thus germane to the issues we have been discussing only insofar as theories which assume that lexical items are decomposed also assume the analyticity (hence, the necessity) of the statements that decomposition gives rise to. It is possible to imagine a kind of semantic theory which makes the first assumption but not the second. Indeed, many of the decompositionalist theories of the lexicon that psychologists have proposed appear to be of that kind; the claim is that they provide a warranted account of the internal representation of lexical items rather than a theory of modal statements.
John is a bachelor or John is a fool. According to A–C, one can understand the sentence without, as it were, recovering the fact that it has this entailment.7 That is, having recovered the representation of *John is a bachelor* as ‘P’, one must then resort to the logic to determine that ‘P’ entails ‘P or Q’. Thus A–C constitute a hypothesis about the dividing line between sentence comprehension and the inferential processes that are defined for sentences. This is because A–C make the claim that sentences containing definable expressions are ipso facto not in normal form for the application of rules of inference; normalizing them requires replacing the definable expressions by their definitions. Any theory which translates sentences into a normal form to which inference rules apply thereby draws a line between comprehension and inference; comprehension is identified with recovering the normal form and inference with operating on it.8

There is, however, no a priori reason for drawing the line between comprehension and inference at the point that A–C specify rather than at some other point. In principle, we could have a more abstract notion of normal form (e.g. one at which not not P and P are identically represented in the output of the comprehension system). Or we could have a less abstract notion of normal form (e.g. one in which the normal form for a sentence respects its surface vocabulary; viz. in which definitions do not replace definables). Of course, compensating adjustments of the inferential system would be required in either case. So, if not not P and P have identical semantic representations, then, to that extent, the inferential system need contain no rule of double negation. On the other hand, if bachelor and unmarried man have distinct semantic representations, then (if we consider bachelor → unmarried man to be valid) some principle to the effect that x is a bachelor only if x is an unmarried man will have to appear among the rules of inference. Such principles are traditionally called meaning postulates.

To summarize: we have given arguments that there is no semantic level in the sense of conditions A–C; in particular, our proposal is that to each morpheme of the surface vocabulary of a natural language there corresponds a primitive expression in the vocabulary of the representational system. (It should be clear that this proposal does not require that the vocabulary of semantic representations is identical with that of any natural language; we do not suppose, for example, that the formatives of the semantic level are phonologically interpreted. All that we require is that formatives of the natural language should correspond to formatives in the representational system, whatever these latter may turn out to be.) If our arguments are sound, then it appears practically mandatory to assume that meaning postulates mediate whatever entail-
ment relations between sentences turn upon their lexical content. That is, meaning postulates do what definitions have been supposed to do by theories which endorse A–C.9

There are, we believe, several further considerations which support the meaning postulate approach. We shall mention just two of them here.

First, notice that the distinction between processes that are involved in understanding a sentence and processes that are involved in drawing inferences from it corresponds to a distinction between mandatory, on-line psychological processes and optional, long-term psychological processes. For, by hypothesis, the output of the sentence comprehension system is that representation of the sentence which must be recovered by anyone who understands it. But the application of principles of inference is presumably largely context-determined. What inferences we draw from what we hear must be a question of what we take to be relevant to the task at hand. (Clearly, we cannot draw all the inferences since there are typically infinitely many.)

Now, the overwhelmingly puzzling problem about sentence comprehension is how people manage to do it so fast. For example, there is evidence that subjects can perform tasks which depend on recognizing meaning relations in sentences with latencies as low as 250 msec; that is, with latencies which approximate the length of a CV syllable or the lower bound on a two-choice reaction task (see Marslin-Wilson 1973). Given this consideration, it seems clear that, barring decisive evidence to the contrary, we should assume that the semantic representation of a sentence is as much like the surface form of the sentence as we can. For, in doing so, we reduce the load on processes that must be assumed to be performed on-line. In particular, then, given a choice between assigning a process to the comprehension system and assigning it to the inferential system, all other things being equal we should choose the latter option. That is precisely what hypothesizing meaning postulates in place of eliminative definitions permits us to do.

Second, it seems clear that even if we do have definitions in our grammar, meaning postulates are still indispensable. For, as has been widely recognized, there will almost certainly be residual inferences that turn on "content" words even after the process of definitional decomposition has gone through. So, for example, even if we have replaced kill by 'cause to die', we will still need some apparatus to mediate the inference from 'x cause y to die' to, say, 'y die'. Such a principle would have to be sensitive to the meaning of 'cause' (notice that 'x intend y to die' does not entail 'y die') and would thus have precisely the character of a meaning postulate.

There is at least one further argument which suggests the indispensability of

9 It is sometimes alleged that theories which employ meaning postulates are just "notational variants" of theories that use definitions (see Katz and Nagel 1974). But, as we have seen, such theories disagree on at least the following: (a) the primitive vocabulary of the language in which semantic representations are couched; (b) the abstractness of the semantic level; (c) the relation between comprehension and inference; and (d) a host of psychological consequences, two of which were examined above.
meaning postulates even in semantic theories that acknowledge eliminative definitions. Some of the recent skepticism about analytic definitions (though by no means all of it) has concerned the possibility of giving logically sufficient conditions for something being $F$, even though the possibility of giving logically necessary conditions for being $F$ is not called into question. In traditional terms, it is often easier to specify the genus than the differentia. For example, many who admit that, necessarily, every case of killing is a case of causing to die, nevertheless deny that every case of causing to die is a case of killing. Nor is it obvious that there is any property of an action (except, of course, being a killing) such that, necessarily, an act of causing to die which has that property constitutes a killing. The problem may be clearer if one considers the relation between red and colored. Presumably $x$ is red entails $x$ is colored. But, surely, there is no property $F$ which is logically independent of the property of being colored and such that $x$ is $F$ and $x$ is colored entails $x$ is red. If this is correct, and if it is analytic that whatever is red is colored, then there is at least one analyticity which does not rest on a definition. For definitions are required to be symmetrical (i.e. they provide sufficient as well as necessary conditions for the application of a term). Meaning postulates, however, are not so constrained. They would thus appear to be the only available mechanism in the present case. (For an elaboration of this argument see J. D. Fodor, forthcoming.)

In fact, all the theories which embrace A–C that have so far been proposed do employ meaning postulates under one guise or other. (In Katz's work, for example, they appear as "redundancy rules".) Since any entailment that can be captured with meaning postulates and definitions can equally be represented with the former alone, it seems clear that the least hypothesis is that definitions are otiose.

We close this discussion by briefly considering three lines of argument which might be held to substantiate the claim that lexical items are decomposed at the semantic level even in face of the counterevidence that we have been assembling.

1. There are claimed to be "purely linguistic" considerations which militate in favor of assigning structure to lexical items.

For example, there have been arguments for lexical decomposition based on the search for a principled distinction between "possible" and "impossible" words. There have been arguments based on the claim that adverbs can have scope over internal semantic components of words. And there have been arguments based on similarities of distribution between words and their defining phrases. Many of these arguments are ingenious and deserve serious examination. It is our belief that there are alternative explanations of the data which do not assume semantic decomposition of lexical items and which are at least as plausible as the ones that do, but establishing this is well beyond the scope of the present article. We must make do with one brief comment on the argument from distributional symmetry and one remark on a matter of general principle.

The distributional arguments we have in mind are those which appeal to the
existence of selectional symmetries between items that are similar in meaning to support the claim that such items must have identical representations somewhere in the grammar. Our point is that such arguments beg the question that is primarily at issue in this article. For, it simply cannot be taken as a self-evident methodological principle that selectional symmetry is necessarily the product of identity of underlying linguistic representation. On the contrary, there is an empirical claim at issue, and what we have been arguing implies that certain distributional symmetries—viz. those between words and their defining phrases—should not be so treated. Nor is it hard to think of alternatives. For example, McCawley (1971) has argued that sentences which violate selection restrictions are simply special cases of sentences with incompatible entailments. If this is so, then it seems that any theory which correctly determines the entailments of sentences could accommodate cases where lexical items and their defining phrases are subject to identical selection restrictions. In particular, a theory which determines entailments by appeal to meaning postulates should be able to do so.

The general point of principle is this. Suppose there were some widely accepted pattern of purely linguistic argumentation that entailed the existence of eliminative definitions. Then if, as we have maintained, eliminative definitions are not psychologically real, that would constitute a reductio ad absurdum of the pattern of linguistic argument, which would then have to be abandoned. This is simply to endorse Chomsky’s insight that the metatheoretical principles that are used to evaluate alternative grammatical arguments and analyses must themselves be empirically justified. In particular, they must be chosen so that they select the correct grammar for each language, viz. the grammar whose constructs are psychologically real.

2. It is claimed that there are fundamental semantic properties and relations which the grammar is required to mark, and the only way to mark them is by assuming a classical semantic level over whose structures they can be formally defined.

The following discussion will suggest the sort of point that is at issue. It is sometimes argued that there is a principled difference between sentential synonymy and logical equivalence of sentences. The suggestion is that this difference can be reconstructed only on the assumption that sentences are synonymous if and only if they receive identical representations at the semantic level. Logically equivalent sentences are then the ones which are interconvertible by valid rules of inference. (A similar contrast is supposed to exist between analyticity on the one hand and (mere) logical necessity on the other.)

To begin with, however, it is worth remembering that the existence of a principled distinction between analytic and logical equivalence has been repeatedly challenged in the philosophical literature. Certainly it is hard to argue that intuitions are clear enough to justify making that distinction correspond to a distinction of levels in the grammar.

But, in any event, there is no reason to suppose that the desired distinction cannot
be drawn within the logic: viz. without reference to a semantic level. For example, one can imagine that intuitions of analyticity might correspond to arguments which employ only a designated subset of the inferential rules (including, of course, the meaning postulates). It is true that, on this view, one might have to list, for each such pattern of inference, whether it preserves analyticity. This may not seem to have much explanatory value, but notice that we do essentially the same thing when we list the dictionary entries and say that analyticity is preserved by applying them.

A different theory of analyticity might have it that intuitions of analyticity are essentially intuitions about the length of arguments under some canonical formalization. Roughly, the longer the argument from $P$ to $Q$, the less analytic the sentence if $P$ then $Q$. If this is true, it would argue for the meaning postulate treatment against a radical division of analytic from logical truth. At the very least, the fact that judgments of analyticity appear to be graded suggests that this might be right.

3. There is a class of psychological effects whose explanation has been held to require the kind of semantic cross-classification of the lexicon which decomposition of lexical items provides.

The best candidates we know of for such effects are slips of the tongue (see Fromkin 1971, Shattuck 1975), the tip of the tongue phenomenon (see Brown and McNeil 1966), and semantic generalization (see Osgood 1953). Without going into details, the point of each of these examples is the intrusion into a subject’s performance in the experimental task of lexical items semantically related to the intended target item. Such intrusions suggest that the speaker/hearer’s access to the contents of the lexicon may be via some system of semantic features. (Thus, for example, the words one thinks of when one attempts to recover some word that is on the tip of one’s tongue are often semantically related to the desired item. So, the subject trying to remember sextant might say astrolabe instead.)

We think these observations are interesting and pertinent but not decisive. Though the existence of semantic intrusion is certainly compatible with the semantic decomposition of lexical items, it quite clearly does not require that view. Any accessing system which makes semantically similar items “closer” than semantically dissimilar ones would yield the same effect even without embracing lexical decomposition. Among the candidate systems are “networks”, multidimensional spaces where the dimensions are semantically interpreted, “semantic fields”, and, surely, many others that ingenuity might devise. The point to bear in mind is that the assumption that lexical items are semantically decomposed does not follow from the assumption that the lexicon is semantically cross-classified (though, admittedly, if decomposition did occur that would provide a convenient mechanism for cross-classification).

Moreover, it is really quite unclear to what extent these intrusions are semantic in nature, if semantic is construed in the disciplined sense that linguists have in mind, such that semantic representation is intimately concerned with the determination of entailment. Our impression is that at least some of the psychological phenomena turn on
associative connections which do not meet this stringent condition. So, for example, *table* is a likely slip for *chair* and *stool* is not. Yet, from the point of view of pure semantics, the amount of overlap in meaning is greater in the latter case than in the former. The difficulty is, of course, that many semantically related words are also high associates, so factoring out the respective contributions of meaning and association to the observed phenomena is no small matter. In any event, it tells against the view that these phenomena are the consequences of lexical decomposition that they appear to hold for proper names as well as for words which have meanings in the linguist's proprietary sense. Slips of the tongue that involve proper names often blend (or otherwise confuse) them with the names of related people: the name that comes to mind when one is trying to think of Smith's name is often that of someone one thinks of as related to Smith, and the names one confabulates with *Rubenstein* are likely to be the names of pianists. We do not propose a theory of any of these processes, but patently they would not justify the lexical decomposition of names.

In conclusion: almost all of the recent discussion about semantics in linguistics, and much of the recent discussion of semantics in psychology, has concerned the detailed arrangement of representations at the semantic level. It seems to us, however, that there is no very convincing evidence for the existence of such a level, and that there is more than a little reason to believe that no such level does exist. Perhaps the reason that semantic representations have proved to be so elusive is simply that, after all, there aren't any.10

References


---

10 We wish explicitly to exempt from these remarks those sorts of representations which are supposed to capture the "logical form" of sentences (as contrasted with the internal structure of their lexical items). Our arguments are not germane to the claim that there exists a level of representation which formalizes such relations as between quantifiers and the variables they bind, those between relational terms and their arguments, or those between operators and what they operate upon. Nor do our arguments commit us, one way or the other, on the relation of that level of representation (if it exists) to the levels of deep and surface syntactic structure. It seems to us that arguments similar in spirit to the ones that we have given can be brought against some of the standard proposals about what representations of logical form are like (cf. Martin 1975.) However, we have not tried to give such arguments here.
UNREALITY OF SEMANTIC REPRESENTATIONS


J. D. Fodor
Department of Linguistics (U–145)
University of Connecticut
Storrs, Connecticut 06268

J. A. Fodor and M. F. Garrett
Department of Psychology (E–10)
MIT
Cambridge, Massachusetts 02139