Alternative conceptions of semantic theory*

ARNOLD L. GLASS
KEITH J. HOLYOAK

Stanford University

Abstract

It is argued that theories of semantic memory have diverged in a manner that parallels current linguistic controversy concerning the representation of meaning. The feature-comparison model (Smith, Shoben & Rips, 1974) applies the linguistic theory of Lakoff (1972) to predict people's reaction times to verify sentences, while the marker-search model, described here, uses the type of semantic representation outlined by Katz (1972) to explain a similar range of data. The two models are described and the evidence for each is reviewed. Available evidence supports the marker-search model, but disconfirms a major prediction of the feature-comparison model. It is argued that the feature-comparison model is in principle inadequate as a model of semantic representation, unless its conception of semantic components is substantially altered.

Philosophers and linguists have long discussed how the meaning of a word is represented in memory. In psychology, semantic memory research has approached this question by investigating possible mechanisms by which people use their knowledge about words to determine whether sentences are true or false. The dependent variable of major interest has been reaction time (RT) to verify simple propositions, such as *A dog is an animal*.

In a recent review paper, Smith, Shoben, and Rips (1974) have proposed a model, called the *feature-comparison* model, to account for the bulk of the RT differences reported in the semantic memory literature. This model emerges from what they term a 'set-theoretic' tradition in semantic memory research (Meyer, 1970; Schaeffer & Wallace, 1970). It uses a semantic theory outlined by Lakoff (1972) as the basis for a psychological process model. Set-theoretic models of semantic memory have been contrasted with net-

*The ordering of authors is haphazard. This paper has benefited from the extensive suggestions of Gordon H. Bower, Daniel Osherson and an anonymous reviewer. We are especially grateful to our good friends Edward E. Smith, Edward J. Shoben, and Lance J. Rips for the free exchange of data and ideas upon which this paper depended.

This paper was completed while A. Glass held an N.S.F. graduate fellowship and K. Holyoak held a Stanford University fellowship, and it was supported by Grant MH13950-06 from the National Institute of Mental Health to Gordon H. Bower.

work models, which represent the meaning of a word as a mapping between the word and a relational network (Collins & Quillian, 1969). However, the distinction between set-theoretic and network models has never been clearly drawn. Smith et al., describe set-theoretic models as those models in which concepts are represented by sets of elements. But network models can also define concepts componentially, so that at this general level the notations of set theory and of graph structures are largely interchangeable. In order to contrast the two types of models more clearly, we will describe two assumptions that can be used to distinguish set-theoretic from network models.

First, set-theoretic models restrict themselves to a very simple formal representation. Each element in the set representing a concept is treated as an atomic unit. Such formal devices as redundancy rules (Katz, 1972), which would permit one element to dominate (and therefore entail) another are excluded from the representation. Semantic relations are defined in terms of operations such as set inclusion; e.g., it might be assumed that a person can verify that a *dog* is an *animal* by determining that the set of features defining *dog* contains the set of features defining *animal*. In contrast, network models can include the graph-theoretic equivalents of redundancy rules in order to mark entailments, while the equivalents of antonymous n-tuples (Katz, 1972) can be used to mark contradictions between semantic components.

A second assumption, one that is central to the feature-comparison model, is that the relation of category membership is in some sense a matter of *degree*. This assumption is identical to Lakoff's (1972) hypothesis that absolute notions of truth and falsity should be replaced by a continuous truth dimension. This view would represent a sentence such as *A bat is a bird* as having some 'intermediate' truth value. In contrast, semantic relations in a network model are basically all-or-none: A component either dominates another, or it does not; and a component either contradicts another, or it does not. This type of representation therefore naturally leads to absolute rather than continuous notions of truth and falsity, as is advocated by Katz (1972). Under this view, a person might be uncertain as to the truth value of *A bat is a bird*, due to his ignorance or the sentence's ambiguity; but nevertheless, an absolute dichotomy remains between truth and falsity. Clearly, these parallel debates in psychology and linguistics are related to the extent that the goals of the two fields converge. Accordingly, the issues which distinguish these two classes of models have implications for linguistics as well as for psychology.

The intent of the present paper is to analyze the empirical and theoretical justification for these two types of models. The first section of the paper provides a critical review of the major evidence for the feature-comparison model of Smith et al. We are focusing our attack on the feature-comparison model for several reasons. First, it formulates the set-theoretic assumptions that we wish to call into question more clearly than any previous proposal. Second, the model has been given a precise formulation that allows the possibility of disconfirmation. Third, as Smith et al., point out, the feature-comparison model has been more successful in accounting for available verification data than any other set-theoretic model yet proposed. Accordingly, if we can show the Smith et al., model to be inadequate we will be rejecting not simply an arbitrary version of a set-theoretic model, but the most successful one yet devised.

The feature-comparison model invokes two separate mechanisms for verifying a sentence. In Section 2, by contrast, we present a network model that adapts the theory of Katz (1972) to psychological prediction, and which assumes a single basic mechanism for verifying sentences. In this alternative model the single underlying variable that determines both true and false RT is the time required for the person to access information that logically confirms or contradicts the truth of the presented sentences. Section 3 presents data that provides support for our proposal, while disconfirming a critical prediction of the feature-comparison model. The final section of the paper then examines broader theoretical issues concerning the nature of semantic representation that are raised by a comparison of the two models. We shall argue that a set-theoretic representation is in principle inadequate as a model of semantic memory.

1. The feature-comparison model: Review and critique

The feature-comparison model assumes that the meaning of a word is represented by a set of features, and that "some features will be more defining or essential aspects of a word's meaning, while others will be more accidental or characteristic features" (Smith et al., 1974, p. 4). Each feature is thus stored along with a weight indicating its degree of 'definingness' for the concept in question. The feature-comparison model posits two distinct serial stages that are used to verify sentences of the form *An S is a P*. In the first stage, the overall relatedness of the subject and predicate words is assessed in terms of all features (regardless of their definingness weights) of the two categories. If the overall relatedness of the subject and predicate words exceeds an upper criterion, a quick 'true' response is made. If their relatedness falls below a lower criterion, a quick 'false' response is made. Only if the overall relatedness falls between the upper and lower bounds is the second stage executed, resulting in a longer RT. This second stage separates the more defining features from the characteristic ones on the basis of feature weights, and compares only the more defining features of the subject and predicate. A 'true' decision is made in case all the defining features of the predicate are contained in the subject; otherwise, the decision is to respond 'false'. The feature-comparison model does not specify any relationship between overall semantic relatedness and the duration of second-stage processing. This model predicts that for true sentences, as the relatedness between subject and predicate increases, the percentage of quick stage-one 'true' responses will increase, resulting in faster mean RT for more related true sentences. But for false sentences, high relatedness will decrease the percentage of quick stage-one 'false' responses, resulting in slower mean false RT as relatedness increases.

We will consider whether the feature-comparison model's assumption of two serial processing stages is justified. This may be done in light of a test derived from one proposed for stage models by Sternberg (1969); viz., are there two conceptually distinct variables, one derived from stage one and the other derived from stage two, that both affect RT as predicted by the feature-comparison model? Smith et al., identify two variables that might satisfy this criterion for the feature-comparison model: (1) Semantic relatedness, which should affect the outcome of stage one only; and (2) category 'size' (i.e., the number of features that define a particular category), which should affect the duration of stage two only. Let us examine the evidence concerning these variables.

Semantic relatedness

For true sentences the feature-comparison model predicts that high semantic relatedness will result in a greater probability of a correct stage-one response, and hence, lead to relatively fast mean RT. Smith et al., review several studies showing that high relatedness indeed speeds up correct classification of an instance as a member of the test category. However, such evidence is open to an alternative interpretation. The problem is that rated relatedness has proved in every case to be positively correlated with the frequency with which the instance is produced as an association to the category name, as measured by association norms such as those of Battig and Montague (1969) (see Rips, Shoben, & Smith, 1973; Rosch, 1973; Smith, 1967; Smith et al., 1974; Wilkins, 1971). For instance the correlation between relatedness and production frequency (one standard measure of association strength) was 0.85 in the Rips et al., study. While this correlation is consistent with Smith et al.'s claim that production frequency reflects semantic relatedness, other evidence demonstrates that production frequency has an independent effect on RT. This evidence is provided by Experiment I of Smith et al., where the correlation between semantic relatedness and production frequency was only 0.49. In that study, production frequency was clearly a better predictor of RT than were relatedness judgments.

Smith et al., can argue, of course, that production frequency simply measures the underlying conceptual variable of relatedness more accurately than do ratings. However, other empirical results are difficult to reconcile with the notion that production frequency can be identified with relatedness. Loftus (1973) obtained measures of the production frequency (PF) of the category given the instance as a stimulus, as well as of the frequency of the instance given the category as a stimulus. She then varied whether the category or the instance was presented to the subject first in a verification task requiring determination of whether the instance was a member of the category. When the instance preceded the category (e.g., *robin-bird*), the instance-to-category PF determined RT; but when the category preceded the instance (e.g., *bird-robin*), the category-to-instance PF determined RT. In terms of the feature-comparison model, this result implies that for the same two words, relatedness differs depending on the presentation order. However, the model does not specify the exact composition rule that is to be used to compute overall similarity, and it is unclear whether or how such a rule can be made sensitive to word order. In other words, if production frequency is to be taken as a measure of relatedness, then the notion of relatedness will have to be considerably complicated.

Other conceptual problems arise in trying to identify production frequency with relatedness. When a person is asked to rate the relatedness of a pair of words, he is being asked to do what Smith et al., assume is done during stage one of sentence verification — i.e., to compare the subject and predicate words and assess their degree of relatedness. Production tasks, on the other hand, involve the retrieval of one concept given another as a cue. While Smith et al., may assume that these tasks measure relatedness, there is no *a priori* reason to believe this to be true. In section 3 we will argue that production frequency reflects a different conceptual variable, namely, the order in which information about word meanings is retrieved. Note that the results of Loftus, described above, have a straightforward interpretation in terms of retrieval. Suppose that verification in the Loftus paradigm requires that the person find a path between the two presented concepts (the category and the instance), beginning at whichever concept is presented first. Then the obtained effects of presentation order simply indicate that the frequency with which the second word is generated as a response to the first in a production task measures how quickly a path can be found from the first concept to the second during verification. Other evidence that production frequency is best conceptualized as a measure of the order of information retrieval is reviewed in Section 3.

Other experimental evidence, reviewed by Smith et al., generally supports their prediction that high relatedness increases false RT. Several studies have found that the RT to reject meaningful (high-related) false sentences (e.g., *All grains are wheats*) is longer than the RT to reject relatively anomalous sentences (e.g., *All typhoons are wheats*) (Kintsch, 1972; Meyer, 1970; Rips et al., 1973; Wilkins, 1971). However, the issue is not yet closed. A study by Glass, Holyoak and O'Dell (1974) suggests that false RT is not monotonically related to overall relatedness of the subject and predicate terms. Contrary to the Smith et al., prediction, false sentences in which the subject and predicate were very closely related (e.g., *Many arrows are dull*) were rejected *more* quickly than relatively meaningful sentences in which the subject and predicate were less related (e.g., *Many arrows are wide*). However, minimally-related anomalous sentences (e.g., *Many arrows are intelligent*) were rejected most rapidly of all. This latter result is not inconsistent with the earlier findings, since previous studies compared a mixture of relatively meaningful sentences, differing in relatedness, to anomalous (very low-related) sentences. Further false RT data that is incompatible with the feature-comparison model is reviewed in section 3 below.

Category size

Smith et al., also specified a variable which supposedly affected only stage two in their model, namely, category size. Since larger, more abstract categories logically have fewer defining features, fewer comparisons should be required in stage two in order to match defining features of a large category with the features of the test instance. Accordingly, holding constant the probability that stage two processing occurs (by controlling semantic relatedness), an increase in category size should decrease decision time. This prediction has been tested by studies that have varied category size while attempting to hold relatedness constant (Landauer & Meyer, 1972; Wilkins, 1971). But contrary to the prediction of the feature-comparison model, both studies found that statements involving larger categories took *longer* to verify than statements involving smaller categories (the difference was 32 msec for false sentences and 17 msec for true sentences, respectively). Smith et al., criticize these studies, arguing that neither study reported tests of the obtained differences against item variability (Clark, 1973). It has not been demonstrated, therefore, that this category-size effect is reliable. However, the fact that these trends are opposite to the prediction of Smith et al., remains problematic.

Finally, Experiment I in Smith et al. tests their category size prediction while escaping methodological problems inherent in the earlier studies. That experiment varied category size and semantic relatedness (as measured by production frequency) independently. The true RT results showed that production frequency was a highly significant variable; but when the variance in RTs attributable to production frequency was eliminated (by an analysis of covariance) the residual effect of category size did not approach statistical significance. Apparently, the conclusion best supported by available results is that, contrary to the prediction of the feature-comparison model, category size in itself has no effect on true RT when production frequency is controlled. Since category size is the sole variable so far proposed to affect stage-two processing, one must conclude that the two-stage model does not meet the evidential standards for multi-stage models proposed by Sternberg (1969).

Smith et al., also test the feature-comparison model by fitting a mathematical model of its major assumptions to the data of their second experiment. The mathematical model provided estimates of such parameters as the length of the duration of stage two in relation to category size, and the subject's criterion (based on relatedness) for making a response without stage two processing. They used two different estimation procedures, one based on relatedness ratings (with sixteen parameters) and one based on error rates (with ten parameters). When RT was predicted from the model on the basis of semantic relatedness ratings the obtained fit was extremely poor, with a correlation between predicted and obtained RT of only $r(14) = 0.69$. Furthermore, the fit provided by the more successful procedure seems to rest on the use of a general correlation between higher error rates and slower RTs. We tested this possibility by predicting RTs for Smith et al.'s data directly from the observed error rates. For this purpose we grouped both the 96 true items and the 96 false items into ten levels of error rates, so as to have at least five items at each level, and used linear regression to predict the mean RTs. For true items the correlation between predicted and observed RT was $r(8) = 0.972, p < 0.01$, and the root mean square deviation equalled 11.3 msec; while for false items the correlation was $r(8) = 0.929, p < 0.01$, with a mean deviation of 16.4 msec. Caution is necessary in

comparing these results with those of Smith et al., since we predicted two sets of ten mean RTs, while Smith et al., predicted a single set of 36 mean RTs. Nevertheless, such a comparison is suggestive. Smith et al., estimated ten parameters in order to predict RT, and obtained a correlation of $r(24) = 0.945$ between predicted and observed RT, with a mean deviation of 28.9 msec. Our predictions, each set of which is based on just two estimated parameters, are no less accurate than those obtained by Smith et al., using their more elaborate model and parameter-estimation procedure. Thus while it is true that the parameter estimates obtained by Smith et al., are consistent with the feature-comparison model*, their RT and error rate data are also consistent with the large class of models that predict a positive correlation between error rates and RT. While this correlation is indeed predicted by the feature-comparison model (see Smith et al., 1974), it is in fact a general empirical result commonly obtained not only in semantic-memory studies but in other RT studies as well (e.g., Clark & Chase, 1972; Meyer & Schvaneveldt, 1971; Posner, 1970). RT and error rate are generally taken to be convergent measures of item difficulty. Consequently, the burden of proof remains with Smith et al., to demonstrate that this relationship reflects processes specific to semantic decision-making, rather than more general response strategies typically used by subjects in RT experiments.**

It should also be noted that the data to which Smith et al., fit their model are drawn from an experiment of a rather problematic design. The subjects' task was to decide whether an instance was a member of a target category (e.g., *bird*), but all distractor

*For instance, Smith et al., found that the parameter estimate for stage-two duration was longer for small categories (280 msec) than for large ones (161 msec), as the feature-comparison model predicts. However, it is possible that this result was artifactual, since category size was confounded with category discriminability in their experiment. For the small categories of *fruit* and *vegetable*, subjects rated the vegetable instances used as more closely related to *fruit* than to *vegetable* (!), while they rated the *fruit* instances used as nearly as close to *vegetable* as to *fruit*. Clearly, *S*s had problems discriminating between what the *E*s called 'true' and 'false' instances for these categories; consequently, *S*s' RTs were slower for these 'small' categories than for statements about instances of the large categories, *animal* and *plant*. However, Smith et al., did not introduce a parameter to account for category discriminability. This difference between the categories therefore had to be reflected by some other parameter. The most likely candidate for this role is the estimate of stage-two duration, since different parameters were estimated for the stage-two duration of large and small categories. Accordingly, the increase in RT for small categories resulting from the difficult discrimination between 'true' and 'false' instances may have been reflected in the parameter estimates by a longer estimate of stage-two duration for small categories. (We thank L. Glass for this suggestion.)

**Smith et al., also found that error RTs are faster than correct RTs, as their model predicts. However, it is not clear whether this is a general effect, as we have found no consistent relationship between error and correct RTs in data of our own. Even if the effect is general, it may be accounted for by a plausible strategy of how response speed might be traded off against accuracy. Subjects may tend to gradually speed up their responses over trials until they make an error. The occurrence of an error may then make the subject momentarily more cautious, and hence slower and more accurate (Rabbitt, 1966). Then he may begin a gradual decrease in RT until the next error. Such a cyclic pattern of response times would result in faster RTs for errors than for correct responses. Furthermore, since difficult sentences require more processing time for a correct decision, and thus are more likely to produce an error if the subject tries to respond quickly, this strategy would also produce a positive correlation between error rate and correct RT across conditions.

instances were drawn from a single non-target category (e.g., *insect*). For this example, the subject could therefore logically decide to respond 'true' not only by verifying that a given instance (e.g., *canary*) was a *bird*, but also by verifying that it was *not* an *insect*. Conversely, he could respond 'false' either after he verified that the instance (e.g., *termite*) was not a *bird*, or after he verified that it was an *insect*. The obtained pattern of RTs suggests that subjects in fact used all these possible decision strategies. True RT was found to depend not only on the relatedness of the instance to the target category (e.g., *canary* to *bird*), as the feature-comparison model would predict, but also on the relatedness of the instance to the non-target category (e.g., *canary* to *insect*). Both of these variables influenced false RT as well. Smith et al., offer only a *post hoc* explanation of these unexpected effects, suggesting that some subjects used different strategies than others, or that subjects varied their strategies from trial to trial (see their Footnote 9). It is clear, however, that these results were not predicted by the feature-comparison model, nor reflected in the parameters of the mathematical model. Nor is it clear that the pattern of results to which Smith et al., fit their model would generalize to a situation in which distractor items were drawn from a variety of categories.

In summary, the feature-comparison model predicts that the time used to make semantic decisions will be determined by two variables: Semantic relatedness and category size. However, these predictions are not unambiguously supported by available data. Production frequency appears to predict true RT more accurately than rated relatedness, while category size does not predict RT at all.

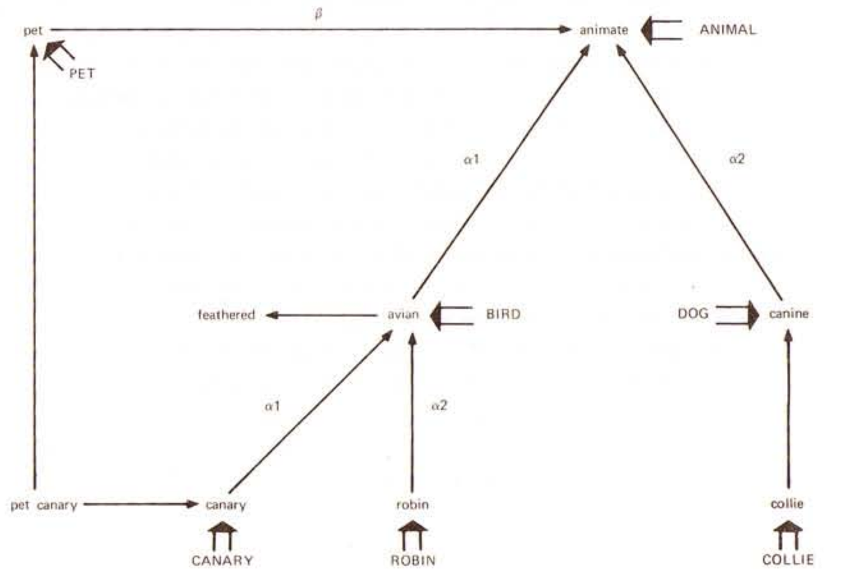
In addition to these empirical difficulties, there are a number of further conceptual problems with the feature-comparison model. A major problem is that the model does not appear to be specified in terms of explicit mechanisms. For example, what mechanisms might plausibly allow a holistic comparison of the type that is postulated to occur in stage one? How are subject and predicate features matched so that a hypothetical overall relatedness estimate can be calculated swiftly? This process must presumably take place without identifying feature dimensions, or else it would seem that a more reasonable and parsimonious strategy would be to compare defining features of the subject and predicate immediately, as is done in stage two of the hypothesized process. Until these questions have been answered, it is impossible to tell how the first stage of the model could ever be executed.

2. An alternative approach: Ordered marker-search

Semantic representation

We will now describe a marker structure (Katz, 1972) in sufficient detail to account for the verification of sentences quantified by *all* or *some* (e.g., *All canaries are birds*, *Some birds are canaries*). These types of sentences are logically equivalent to the cate-

Figure 1. *Hypothetical marker structure, illustrating word-to-marker and marker-to-marker associations.*



gorization tasks to which the feature-comparison model has been explicitly applied (e.g., deciding that *canary* is an instance of the category *bird*, or that the category *bird* contains the instance *canary*). No attempt will be made to describe how relationships other than category membership could be represented, such as possession (*has*) or ability (*can*). Thus the only verb which can be represented in the structure to be described is the present tense of the copula (*is, are*). Discussion of the extendability of the two models will be postponed until the final section of this paper.

A portion of such a marker structure is diagrammed in Figure 1. In this figure labels representing words are in capital letters, while labels of markers are in small letters. Thus BIRD represents a word, while (avian) stands for a marker. Where possible, markers are labeled with adjectives to emphasize that they are best thought of as properties (e.g., (animate)), rather than as categories or exemplars. For many markers no appropriate English adjective with which to label them exists; these are labelled with nouns. Thus the marker labeled (canary) stands for an abstract concept roughly equivalent to "possessing the essential properties of a canary".

As Figure 1 illustrates, we assume that most common words are directly associated with a single marker in the attribute structure. In each case we will refer to this marker as the 'defining' marker for the particular word. For example, (avian) is the defining marker for *bird*. In the case of words with multiple meanings (e.g., *bank*), each sense of the word will access a different defining marker. Note that no single word is associated with the marker labeled (pet canary). This illustrates our assumption that only a subset of the markers in memory are directly associated with words. But we assume that all markers can potentially be accessed during the search procedure used to verify sentences.

The marker-search theory has two basic structural assumptions. First, we assume that markers are interrelated in such a way that any marker stands for, or dominates, a set of further markers associated with it. In Figure 1 these associations are represented by arrows. For example, the arrows pointing from (avian) to (animate) and (feathered) indicate that (avian) stands for the set {(animate), (feathered)}. The property of containment represented by the arrows is transitive. Thus in Figure 1, since the marker (robin) implies – that is, dominates – the marker (avian), it must also dominate (animate) and (feathered). In other words, (robin) stands for the set of markers {(avian), (animate), (feathered)}. Also, by definition any marker dominates itself.

The arrows in Figure 1 are labeled in order to illustrate the second basic structural assumption of the model – that information about contradictions is represented in memory through the associations between markers. In the network structure a contradiction arises whenever two arrows labeled with the same greek letter meet at a common node. In the figure arrows which connect animal species to (animate) are labeled with an α, while the arrow connecting (pet) to (animate) is labeled with a β. Thus a contradiction arises at (animate) between (avian) (α_1) and (canine) (α_2), but not between (avian) and (pet) (β). This notational system simply represents our intuitive knowledge that a bird cannot be a dog, but might possibly be a pet. Note that no significance is attached to the labels used (α, β, etc.). The essential point is simply that intersections are of two types – contradictory and non-contradictory. It should also be noted that in Figure 1 the marker (robin) dominates (avian) while (collie) dominates (canine). It therefore follows from the transitive properties of the arrows that a contradiction of the sentence *A canary is a collie* also arises at (animate).

This treatment of redundancy rules and antonymy is based on the proposals of Katz (1972). The present notation is a variant of that used by Katz, who introduces superscripted markers to form antonymous n-tuples. The notation we are using is similar to that suggested by Bierwisch (1969), and has the advantage of capturing the close relationship between redundancy rules and contradictions.

The different logical relationships between concepts that are denoted by the quantifiers *all* and *some* (*All S are P* and *Some S are P*) can be specified in terms of the relationships between the defining markers of the subject and predicate (see Table 1). An *All*-statement is true if the defining subject marker dominates the defining predicate marker. For example, since in Figure 1 an arrow points from (avian) to (animate) it follows that *All birds are animals* is true. Similarly, since (avian) also implies (feathered), *All birds are feathered* is true as well. The truth conditions for *some* are slightly different. A *Some*-statement is true if there exists a marker *from* which arrows lead to both the defining subject marker and the defining predicate marker. The sentence *Some animals are feathered* is therefore true, since (avian) satisfies the above criterion.

Table 1. *Decision criteria for true and false sentences quantified by* All *or* Some

Quantifier	True sentences: Criterion	True sentences: Examples	False sentences: Criterion	False sentences: Examples
All	Defining subject marker dominates defining predicate marker	All birds are animals.	Defining subject marker contradicts defining predicate marker	*Contradictory:* All birds are dogs.
			A marker which dominates defining subject marker contradicts defining predicate marker	*Counterexample:* All birds are canaries.
Some	A marker dominates both defining subject marker and defining predicate marker	Some birds are canaries. Some birds are pets.	Defining subject marker contradicts defining predicate marker	*Contradictory:* Some birds are dogs

As discussed above, falsity based on contradiction can also be explicitly defined in terms of the relations among arrows. Both *All-* and *Some*-statements are false if a contradiction exists between the defining subject marker and the defining predicate marker; i.e., if similarly-labeled arrows from the two defining markers meet at a common node. For example, we have already seen that (avian) contradicts (canine); consequently, *All/ Some birds are dogs* is false.

In addition, an *All*-statement can be falsified by a slightly different type of contradiction. An *All*-statement is also false if there exists a marker which dominates the defining subject marker, but contradicts the defining predicate marker. *All birds are canaries* is therefore false because (robin) satisfies the above criterion, and thus serves as a counterexample to the sentence – it stands for a bird that is not a canary. Notice that this type of false *All*-statement would be true if the quantifier were *some* (e.g., *Some birds are canaries*). This kind of false *All*-statement will be referred to as a "Counterexample" sentence. The distinction between Counterexample sentences and those falsified by a direct contradiction (Contradictory sentences) will play a critical role in predicting the time taken to reject false sentences.

An interesting property of this representation is that it allows the person to be uncertain about the truth of certain sentences. For example, the sparse network of Figure 1 does not contain the information that some dogs are pets, since there is no marker equivalent to (pet dog) which dominates both (dog) and (pet). On the other hand, (dog) and (pet) have a non-contradictory intersection at (animate), indicating that a dog *might* be a pet. So if this memory were probed with the sentence *Some dogs are pets*, neither the criteria for a 'true' nor for a 'false' decision could be satisfied, and the appropriate response would be something like "It's possible, but I'm not sure". A real subject might give a similar answer to a sentence about some really obscure fact such as *Some anteaters are pets*.

We should point out that our discussion of structural representation has left unresolved a number of serious problems associated with marker theory. We have nothing to say about what elements in the marker system are primitive, except to agree with Katz (1972) that this question cannot be decided *a priori*, but only after considerable empirical study. Other important issues that have been ignored concern how a marker system could be acquired developmentally. For instance, how are redundancy relations generated between markers, and what determines whether an intersection between two relations will be marked as contradictory? While the evidence we will discuss below is solely from studies of sentence verification by adults, the adequacy of any representational proposal cannot be firmly established without studies of how a system of information can be acquired.

Searching the marker structure

According to Katz (1972), the set of markers that form the dictionary entry for a word is unordered. Accordingly, while the hierarchical structure in Figure 1 serves to represent the logical relations between markers, it is not necessary to assume that it indicates the order in which the markers are accessed to verify a sentence. If this assumption *could* be made, it would provide a strong empirical and formal constraint on network representations (Collins & Quillian, 1969). However, there is now ample evidence that people sometimes access relatively abstract markers more quickly than those that are less abstract. For instance, people can decide that scotch is a drink more quickly than they can decide that it is a liquor (Smith et al., 1974). Examples such as these demonstrate that the order in which markers are accessed is not always hierarchical.

However, the fact that the marker set is unordered in Katz's theory does not mean that the markers are accessed in some random fashion. It is possible to specify a performance model based on ordered, though non-hierarchical, search procedures. This possibility can be realized by modifying the structural representation in Figure 1 to include additional redundancy-rule pathways. For example, we will allow the possibility of a direct link between (canary) and (animate), as well as the illustrated pathway from (canary) to (avian) to (animate). To establish an ordering between any two alternative pathways, we can assume either that one of the pathways has a higher probability of being searched first, or that both pathways are searched in parallel with one requiring less time to traverse.

Postulating additional redundancy rules immediately raises a serious question: How is the model to be constrained? Abandoning a strictly hierarchical representation leaves the model without any *formal* constraint that would prevent each node from being directly connected to every node that it dominates. It would be possible to fit the pattern of RTs from any sentence verification task *post hoc*, simply by adding to the structural representation whatever additional connections are required. The price of this freedom, of course, is that the model would be rendered entirely vacuous.

But while no *formal* constraints on the representation prevents this outcome, it may be possible to find *empirical* constraints that will yield specific, testable predictions. This is the strategy that is followed in the studies to be reviewed below. First an empirical measure of search order, independent of RT, is identified. Since the same representation is presumably used in all tasks that depend on the kind of semantic information we have described, the results from one such task should predict performance on the other. Secondly, verification of different sentences will sometimes depend on the *same* semantic information. The relative speed with which a particular bit of information can be accessed should therefore determine the RT to verify a number of different sentences. In general, the complexity of the representation that will be required to account for sentence verification is an empirical issue. If it eventually becomes necessary to postulate unlimited redundant connections in the network, the marker-search model will be rejected as unworkable. But if sufficiently strong empirical constraints can be maintained, the model will have explanatory value.

The marker-search model takes the order in which markers are searched to be the underlying variable determining differences in semantic decision time. In the studies discussed below, we assume that the search procedure is subject to the following constraints: (1) During sentence verification the markers accessed by a word include not only those markers dominated by the defining marker of the word, but also those which dominate the defining marker. For example, consider the word *bird* in Figure 1, which is defined by the marker (avian). We assume that the marker search-set for the word *bird* will include both (animate) (which is dominated by (avian)), and (canary) (which dominates (avian)). This assumption is necessary to explain how the various types of true and false sentences summarized in Table 1 can be verified. Decisions about true *All*-statements and Contradictory false sentences are based on a marker dominated by the defining subject marker; decisions about true *Some*-statements and Counterexample *All*-statements, on the other hand, are based on a marker dominating the defining subject marker. (2) The order in which the markers accessed by a word are searched is independent of the particular quantifier or the truth value of the sentence in which the word appears. Thus the markers of *building* will be accessed in the same order in sentences such as *Some buildings are houses* and *All buildings are houses*. (3) The search will self-terminate as soon as either the criterion for a 'true' or for a 'false' response (as discussed above) is satisfied. For example, a 'true' response will be made to the sentence *All birds are animals* as soon as (avian) is found to dominate (animate); while a 'false' response will be made to the sentence *All birds are dogs* as soon as (avian) is found to contradict (canine).

It would presumably be possible to construct a number of different explicit search mechanisms consistent with the above constraints. These might differ in features such as the degree to which search is serial or parallel. However, the predictions outlined below will not discriminate between the various possible search models based on the semantic representation and search constraints described.

3. Evidence for the marker-search model

In this section we review the evidence for the search model provided by recent experimental results. The model takes the order in which markers are searched to be the variable that determines differences in semantic decision time. In order to test this model, it was necessary to develop a measure, independent of RT, of the order of marker search. Such a measure was proposed by Glass et al. (1974). They asked subjects to provide true one-word completions for incomplete sentences of the form *All/Some S are*________, and tabulated the frequency with which different words were given as predicates. This constrained association technique is similar to the way production frequency norms are collected from subjects who are asked to produce different instances as responses to a category name (Battig & Montague, 1969). Glass et al., assumed that the frequency with which a word appeared as a completion reflected the probability with which its corresponding defining marker was accessed from the defining subject marker.

Clearly production frequency can be at best an imperfect measure of search order, since we have assumed that many markers will not correspond to single common English words. This problem is particularly acute in the case of anomalous false sentences, in which the subject and predicate words generally differ at the level of abstract markers such as (living) versus (non-living) (e.g., *All birds are chairs*). These sentences are typically rejected relatively quickly (Kintsch, 1972; Wilkins, 1971); but since the abstract markers on the basis of which they could be rejected seldom define common English words, production frequency is not a valid measure of the speed with which such markers are accessed. The problem of predicting RT to reject anomalous sentences is discussed more fully in Holyoak and Glass (1975). In the present paper we will only discuss sentences for which production-frequency measures make clear RT predictions; i.e., true sentences, Counterexample sentences, and Contradictory sentences in which the subject and predicate words differ with respect to relatively specific markers (e.g., *All birds are reptiles*).

The marker-search model predicts that true sentences with high production frequency (PF) will be verified more quickly than sentences with lower PF. Glass et al. tested this prediction for sentences with five different quantifiers (*All, Many, Some, Few* and *No*), and both noun and modifier predicates (e.g., *All birds are animals, All birds are winged*). In each case the corresponding PF norms successfully predicted RT. These results extended the findings of Loftus (1973) and Wilkins (1971), who also found that high PF leads to fast true verification RT. The Glass et al., findings have since been replicated by Glass and Holyoak (1974) and Holyoak and Glass (1975). We have also seen that in studies where semantic relatedness was a successful predictor of true RT, it was confounded with PF (Rips et al., 1972; Rosch, 1973). Furthermore, in the one case in which the effects of the two variables have been compared, PF was a much better predictor than was relatedness (Smith et al., 1974). Accordingly, the available data concerning true RT are consistent with the model.

Generation of false sentence completions

While previous experiments were primarily concerned with using PF (or relatedness) to predict true RT, the results obtained by Glass et al., for false *Many*-statements (discussed earlier) indicated that it should also be possible to use such norms to predict RT to reject false sentences. These predictions, to be outlined below, were tested by Holyoak and Glass (1975).

As a necessary initial step, Holyoak and Glass collected false PF norms for sentences quantified by *all* or *some*. They had 32 Stanford undergraduates generate *false* completions for sentences of the form *All S are*________ and *Some S are*________. The resulting false PF norms were compared with true PF norms compiled in previous work (Glass & Holyoak, 1974; Glass et al., 1974). Several striking relationships between true and false sentence completions provided evidence for the marker-search model. Referring to Figure 1, let us consider strategies that people might use to generate false completions of a sentence. One plausible strategy would be to first access a marker dominated by the defining subject marker, and then use it to compute a contradiction as a response. For example, when presented with *All/Some birds are______*, the person might follow the arrows from (avian) to (animate), and then from (animate) to (canine). This procedure would generate the false completion *dog*, producing a Contradictory sentence, as defined in Table 1. If this strategy were actually used by subjects, and search order is independent of the quantifier (as the model assumes), there should be a close relationship between the frequency with which *All/Some birds are dogs* is produced as a false sentence, and the frequency with which *All birds are animals* is produced as a true sentence. Specifically, we expected that each high-PF true completion of *All S are________* from the norms of Glass and Holyoak (1974) and Glass et al (1974), would determine some high-PF Contradictory completion of both *All S Are________* and *Some S are________*. As the examples given in the top of Table 2 illustrate, this prediction was confirmed. Fourteen of the 16 highest frequency true *All*-statements (produced by between 35% and 78% of the respondents) corresponded to high-PF Contradictory sentences (produced by from 19% to 56% of respondents).

Referring again to Figure 1, let us illustrate a second strategy that subjects might use to generate false *All*-statements. This strategy depends on our assumption that markers that dominate the defining subject marker are also accessed by the subject word. The subject can therefore go directly from the defining subject marker to a marker that dominates it, and then use this marker to produce a false completion. For instance, given the fragment *All birds are__________*, the person can simply follow the arrow from the (avian) to the (canary) marker, and respond with the word *canary*. This procedure

Table 2. *Examples of relationships between high-frequency true and false sentences*

Contradictory falses	
True	False
All birds are animals.	All/Some birds are dogs.
All chairs are furniture.	All/Some chairs are tables.
All women are humans/females.	All/Some women are males.
All diamonds are stones.	All/Some diamonds are emeralds.
Counterexample falses	
True	False
Some flowers are roses.	All flowers are roses.
Some prisoners are men.	All prisoners are men.
Some books are novels.	All books are novels.
Some teachers are professors.	All teachers are professors.

would produce Counterexample sentences, such as those shown in Table 2. Note that exactly the same procedure would produce a *true* sentence if the quantifier were *some* (e.g., *Some birds are canaries*). Since the model assumes that search order is independent of the quantifier, our second prediction was that each high-PF true *Some*-statement (from the earlier norms of Glass and Holyoak and Glass et al.) would correspond to a high-PF Counterexample *All*-statement. This prediction was confirmed. The bottom of Table 2 lists four examples of the 22 highest frequency true *Some*-statements from the earlier norms (given by from 22% to 89% of respondents). Each of these 22 true *Some*-statements (e.g., *Some flowers are roses*) corresponded to a high-frequency false Counterexample sentence (e.g., *All flowers are roses*), given by from 16% to 53% of respondents.

A comparison of true and false sentence completions therefore supported the assumption that contradictions can be found in memory. People appear to access the same markers in producing false as well as true sentence completions, except that they contradict the marker in generating false completions.

Rejection of false sentences

The central evidence that discriminates between the marker search and feature-comparison models concerns RT to reject false sentences (Holyoak & Glass, in preparation). The marker-search model predicts that disconfirmation of meaningful false sentences (i.e., Contradictory and Counterexample sentences) requires discovery of a contradiction. Consequently, the sooner the person can access a marker that brings out a contradiction between the subject and predicate, the quicker such sentences will be rejected. For both Contradictory and Counterexample sentences, the order in which contradictions are discovered should be predicted by our production frequency norms; however, the variable that determines false RT should be quite different for these two kinds of false sentences.

Contradictory sentences (e.g., *All/Some birds are dogs*) contain predicates that directly contradict the subject (See Table 1). For such a sentence, its production frequency was taken as an index of the speed of accessing the marker (e.g., (animate)) that produces a contradiction between the subject and predicate. For Contradictory sentences, then, false statements that have high frequency in the norms should be rejected more quickly than false statements given with low frequency in the norms. In agreement with this prediction, Holyoak and Glass found that high-PF Contradictory sentences quantified by *all* or *some* were rejected significantly more quickly than low-PF Contradictory sentences (1319 versus 1468 msec). This result extended the previous findings for false *Many*-statements obtained by Glass et al. (1974). Furthermore, Holyoak and Glass had subjects rate the relatedness of the subject and predicate words for each of their false sentences, and found that their high-PF sentences were rated as significantly more related than their low-PF sentences. Since the feature-comparison model of Smith et al., (1974) predicts that high-related false sentences will be rejected *slower* than less related false sentences, these RT results are opposite to the prediction of the feature-comparison model.

In Counterexample sentences (e.g., *All birds are robins*), the predicate does not directly contradict the subject. The marker-search model therefore predicts that in order to reject this type of sentence, the person must discover some marker (representing an exemplar) that dominates the defining subject marker (e.g., (canary)) and that contradicts the predicate. Accordingly, the RT should be fastest for those sentences for which a disconfirming counterexample was produced most frequently as a true *Some*-completion. Specifically, the RT to reject a Counterexample sentence such as *All birds are robins* should be faster the higher the frequency with which the most common counterexample (e.g., 'canary') was given as a true completion of the sentence *Some birds are*________. Since there is no direct contradiction between the subject and predicate for this type of sentence, the production frequency of the sentence itself (*bird* to *robin*) should have no appreciable effect upon the time to reject it.

To test this prediction, Holyoak and Glass selected Counterexample sentences in which the PF of the most common counterexample was varied orthogonally with the PF of the sentence itself. As predicted, sentences with high-PF counterexamples were rejected significantly more quickly than sentences with low-PF counterexamples (1397 versus 1506 msec), while the PF of the sentence itself had no significant effect on RT. Since PF and semantic relatedness were again positively correlated, the feature-comparison model is unable to account for these results.

To summarize, experimental evidence supporting the marker-search model comes from several sources. The model accounts for observed semantic relationships between true and false sentence completions. It predicts the strong correlation between production frequency and RT to verify true sentences. Most strikingly, the model successfully predicts the RT to reject meaningful false sentences (both Contradictory and Counterexample sentences). The same data for false sentences disconfirm a major prediction of the feature-comparison model, namely, that false sentences with subject and predicate words closely related in meaning are necessarily slow to be rejected.

4. Issues in semantic representation

In this section we examine some theoretical issues that can be highlighted by a comparison of the feature-comparison and marker-search models. These issues center on the distinctions between set-theoretic and network models outlined earlier, and particularly on the differing conceptions of the representation of word meaning which emerge from the two approaches.

Semantic redundancy rules and category size

The marker-search model postulates abstract markers representing the entire set of elementary components associated with the definition of a word (e.g., the marker (avian) would dominate all the markers that define the word *bird*). The markers that define the subject and predicate can then be matched directly during sentence verification; e.g., to verify *All canaries are birds* the system need only discover that a set inclusion relation holds between (canary) and (avian). But should it be necessary, more elementary components can also be recovered by means of associations representing semantic redundancy rules (Bierwisch, 1969; Katz, 1972). Thus, since a set-inclusion relation holds between (canary) and (avian), and also between (avian) and (animate), the sentence *All canaries are animals* can also be verified. While the latter sentence requires a longer search, the final match that allows a decision is again between just two markers ((canary) and (animate)). Accordingly, the 'size' of a category, in terms of the number of elementary components associated with the defining marker of a word, is not a relevant variable in predicting RT.

In contrast, the feature-comparison model does not postulate markers that represent sets of meaning components. The features of such a model are not structured hierarchically. This theoretical distinction yields a very different description of the verification process than that given by the marker-search model. In the second stage of the feature-comparison model, verification of *All canaries are birds* would require matching the set of defining features of *bird* (e.g., 'feathered', 'egg-laying', 'breathing', 'solid', etc.) with the defining features of *canary*. If the predicate were *animal*, its feature list would be shorter, so that second-stage processing should require less time. But as indicated in the first section, there is no evidence that this prediction holds. The introduction of redundancy rules as a psychological construct thus serves at least two functions: It simplifies the search process presumed to occur during verification, and it explains the negative finding that sentences with relatively abstract predicate categories are not verified any more quickly than sentences with less abstract predicates, as long as production frequency is controlled.

The defining/characteristic distinction

A related issue concerns Smith et al.'s proposal that features can be weighted by their degree of 'definingness', ranging from clearly defining to simply characteristic features. They cite Lakoff's (1972) analysis of hedges as major linguistic support for this defining versus characteristic distinction among features. In particular, three hedges are alleged to clearly differentiate the types of features shared by the subject and predicate of certain sentences. According to their analysis, the sentence *A robin is a true bird* is acceptable because *robin* shares both the defining and characteristic features of *bird*, since the features characteristic of a category are those that define common instances. The sentence *Technically speaking, a chicken is a bird* is acceptable because *chicken* shares the defining but not the characteristic features of *bird*. Finally, the sentence *Loosely speaking, a bat is a bird* is acceptable because *bat* shares the characteristic features of *bird*, but not the defining ones. If the hedge in any of these examples is replaced with one of the other hedges (e.g., *Technically speaking, a bat is a bird*), the resulting sentence is less acceptable, since the subject and predicate do not share the type of features specified by the hedge. Smith et al., conclude that this linguistic result is best explained by assuming that features differ in the degree to which they define words.

However, one may explain such hedges in another way. It seems that certain common English words have at least two definitions. One is a popular definition, often learned early in life, such as the fact that a bird is a small flying animal with wings. The other is a technical definition, first agreed upon for some specific purpose by scientists or lawyers, and eventually picked up by dictionary writers and imposed upon the general public. The biological definition of a bird is an example of a technical definition. When an instance fulfills the requirements of both the technical and popular definitions, it may be said to be 'a true' member of the category (as *robin* is for *bird*). When an instance fulfills only the technical definition (e.g., *chicken* for *bird*), then we say 'technically speaking'; when only the popular definition is satisfied (e.g., *bat* for *bird*), we say 'loosely speaking'.

This explanation of hedges can account for the unacceptability of certain sentences that create problems for Smith et al.'s defining/characteristic explanation. For instance, if we use a category as the subject of a sentence, and a 'true' instance as the predicate, the subject will still share characteristic features with the predicate. Accordingly, the hedge 'loosely speaking' should apply, producing such examples as *Loosely speaking, a bird is a robin*. But such sentences are unacceptable. A similar difficulty occurs with two 'true' instances of a category, which surely share many characteristic features. The Smith et al., analysis therefore predicts that a sentence such as *Loosely speaking, a robin is a canary* should be acceptable, but again it is not. The defining/characteristic explanation of hedges requires some additional assumption to explain these cases.* In contrast, the unacceptability of these latter sentences follows directly from our popular/technical explanation, since in neither of these cases does the predicate appear to have two definitions, nor does the subject satisfy any definition of the predicate.

Furthermore, one suspects that the as yet unspecified characteristic and defining features on which the hedges 'technically' and 'loosely' are supposedly based can never be specified. Do people know that the sentence *Technically speaking, a whale is a mammal* is true because a defining feature for *mammal* is shared by *whale*? Or do they know it is true simply because they know that there exists a technical definition of *mammal* that includes *whale*, even though they don't know what it is? In this case, the technical definition actually becomes definition by enumeration.

*One such additional assumption would be that the predicate category must always be more general than the subject category. However, it is not clear that an adequate metric of generality can be specified in terms of the feature-comparison model. One might suppose that more general categories are those with fewer defining features. But this tack allows comparisons of generality only when the categories are logically nested. But two instances, such as *canary* and *robin*, are not nested; consequently, the 'number of defining features' metric does not specify how we know that *canary* is not more general than *robin*. There is no apparent *a priori* reason to suppose that all instances have an equal number of defining features.

An explanation of these hedges in terms of popular and technical definitions has certain testable implications. For instance, we would not expect people in primitive 'non-technical' cultures to make a distinction equivalent to the difference between 'loosely speaking' and 'technically speaking'. Also, since children presumably acquire popular definitions of words earlier than technical definitions, young children should classify an instance as a 'correct' member of a category if and only if it satisfies the popular definition. For example, young children presumably would classify a bat as a bird, and a whale as a fish.

The status of network models

Smith et al., cite the evidence from hedges as a general source of difficulty for network models of semantic memory, of which the marker-search model is an example. They do point out that a network model could incorporate the notion of popular and technical definitions of words by including separate markers for the two definitions. But Smith et al., object to this tack as unparsimonious, presumably because it leads to a proliferation of markers. But it is not clear that this solution is less parsimonious than that offered by the feature-comparison model, which is to assume that a weight indicating 'definingness' is stored with every feature-category pair in memory. That is, given a model based on meaning components, is it more parsimonious to add a finite number of extra components (e.g., popular and technical definitions) or a completely new theoretical mechanism (e.g., definingness weights)?

We should make clear that every use of a hedge does not require a network model to postulate an additional definition for a word. Thus Smith et al., suggest that the acceptability of the sentence *A decoy is a fake bird* requires a network model to postulate an additional definition of *bird* representing 'pseudo-instances'. But in terms of the semantic theory of Katz (1972), the definition of *fake* would be represented by markers corresponding to 'intended to appear as'. Through syntactic and semantic amalgamation rules, a *fake bird* would thus be an object intended to appear as a bird. Any word with defining markers that contain the markers of this compound (such as *decoy*) would be correctly classified as a *fake bird*. This treatment of the meaning of a noun plus modifier could be incorporated into a network model. In contrast, Smith, Rips and Shoben (1974) claim that *fake* is used when the subject and predicate share only characteristic features. However, this proposal is clearly incorrect, as is demonstrated by the unacceptability of the sentence *A bat is a fake bird*. While a bat, just like a decoy, both looks like a bird and isn't a bird, its resemblance is essentially accidental rather than intentional. Consequently, this sentence is unacceptable. It is not clear, however, how the feature-comparison model could distinguish between the acceptability of these two examples simply on the basis of feature overlap.

A second criticism that Smith et al., direct at network models is the finding that typical instances are categorized more quickly than atypical ones (Rips et al., 1973;

Rosch, 1973; Smith et al., 1974). Smith et al., discuss two ways in which network models could incorporate this finding. It may be that intermediate nodes are interposed between the markers representing an atypical instance and the category (e.g., (chicken) might first imply (domestic bird), which then in turn implies (avian)). Also, a network model could account for the effect of typicality on RT simply by assuming that the defining marker of the category is higher on the search list for typical instances than for atypical instances.

Both these possibilities seem plausible. As Smith et al., point out, typicality effects pose difficulties only for those network models that assume that search order somehow mirrors the logical structure of the concepts (Collins & Quillian, 1969). But since the marker-search model does not make this assumption, it avoids these difficulties. To illustrate, consider how the marker-search model handles one typicality result obtained by Smith et al., which they view as particularly troublesome for network models. To use their example, they found that *robin* is more typical than *chicken* for the superordinate *bird*, whereas *chicken* is more typical than *robin* for the superordinate *animal*; moreover, this interaction is reflected in verification RT. This result can be represented very simply in terms of the marker-search model. The markers (chicken) and (robin) can both have direct associations to both (avian) and (animate). If search is serial, in each case there must be a trade-off in terms of whether the association to (avian) or to (animate) is searched first. Which ever has priority, the other must be accessed more slowly. Production frequency norms would presumably indicate that for (chicken) the association to (animate) is searched relatively early, whereas for (robin) the association to (avian) has priority. Consequently, people will be relatively quick to verify both *A chicken is an animal* and *A robin is a bird*, but slow to verify *A chicken is a bird* and *A robin is an animal*. Not only is this result consistent with the marker-search model, it in fact provides evidence for the assumption that search is reliably ordered.

The basic issue, therefore, is whether the effects of typicality are best explained by a model based on continuous variation in degree of category membership, or by a model based on the discrete nodes of a network structure. The linguistic and experimental results cited by Smith et al., do not discriminate between these two conceptions of semantic memory.

Can the feature-comparison model be extended?

In order to handle the RT results reviewed in the previous section, the feature-comparison model could incorporate ordered search strategies of the type we have described as an elaboration of the second processing stage. 'False' decisions could be based on the discovery of a contradiction, defined in terms of a relationship between features. But it would then appear that the first stage of the revised model would either be superfluous or redundant, unless RT data can be found that could not be explained by this new second-stage process alone. If the processing mechanisms of the feature-comparison model were to become more clearly specified, it might therefore become more similar to a network model.

It is important to evaluate current theories not only in terms of how well they handle available data, but with respect to their prospects for extension to the wider range of phenomena with which cognitive psychology must eventually deal. So far semantic-memory research has been almost exclusively concerned with meaning-comparison tasks involving just two words, sometimes with variations in the quantifier. The feature-comparison model can account for some of this data reasonably well. But if we take as our goal the description of the psychological representation of meaning, our theories must eventually analyze verbs, and allow us to represent the meaning of sentences of considerable complexity. Will it be possible to extend the feature-comparison model to accomplish this task?

There is reason to doubt it. To understand the apparent theoretical deficiencies of the feature-comparison model, it is helpful to distinguish between two aspects of meaning, which in linguistics are termed 'reference' and 'sense' (Frege, 1952). 'Reference' refers to the relationships between words and objects or events in the world, while 'sense' refers to the relationships of words to other words – that is, to their meanings within the linguistic system. Both of these aspects of meaning represent important problems for psychology, and in many respects they are clearly interrelated (e.g., in regard to the acquisition of word meanings). Nevertheless, the conceptual distinction between reference and sense is an important one for theories of semantic memory. The intuition underlying the feature-comparison model – that some instances are more 'typical' members of a category than others – essentially concerns referential meaning. However, the verification tasks that have provided the data base for semantic-memory models require the subject to rapidly compare word meanings – that is, to evaluate the 'sense relations' between words. The main structural assumption of the feature-comparison model, that word meanings are represented by sets of features weighted on their definingness, constitutes a hypothesis concerning the representation of sense relations.

The feature-comparison model has thus been directed at the problems of both reference and sense, although the distinction has not been kept clear. The marker-search model, on the other hand, is directly concerned only with sense relations. In the framework of the latter model, a major theoretical problem is to specify how a system of semantic markers can be mapped onto the perceptual system to account for people's basic ability to use words to refer to objects and events in the world. However, in the present paper we are solely concerned with evaluating semantic-memory models as accounts of the representation of sense relations. We will argue that the feature-comparison model lacks the theoretical power to represent a variety of concepts.

As a set-theoretic model, the feature-comparison model treats features as if they were independent 'units' of meaning. The decision process in both of the two stages in the feature-comparison model can be conceptualized as the 'summing up' of information from a set of independent comparisons between pairs of features. The principal theoretical device of this model – the weighting of features with respect to their definingness – thus makes it disturbingly similar to a class of weighted-feature models (perceptrons) that Minsky and Papert (1969) have proved to be in principle inadequate as theories of pattern recognition. The basic problem with such models is that they are unable to recognize visual properties such as connectedness, which depend on the relationships between features. If concepts in natural language can also be shown to be relational, this would suggest that perception-type models will also prove inadequate in the domain of language.

With respect to the representation of sense relations, a set-theoretic feature theory faces two hurdles – first, the problem of specifying a set of independent features; and second, the necessity to demonstrate that these features are sufficient to define concepts in natural language. It seems possible that these hurdles will prove insurmountable. A simple set-theoretic model can represent the relation of containment, as expressed in a sentence such as *A canary is a bird*, by assuming that the set of features defining the predicate is a subset of the features defining the subject. But how could it represent the relation of possession, which is expressed by a verb such as *has* (e.g., *The man has a car*)? This sentence does not refer to an entity that combines the features of *man* and *car*; nor does the sentence imply that *man* and *car* share some features. It does not seem possible to represent the abstract relationship expressed by *has* (or many other verbs, such as *can* and *does*) by combining sets of features using the simple boolean operations to which set-theoretic models have restricted themselves.

If the relation of possession cannot be represented by a set-theoretic feature model, then such models will also be unable to account for category membership, the problem which the feature-comparison model allegedly addressed directly. For example, instances of *money* are defined by their function as a medium of exchange. If we then analyze the concept 'exchange', we find it refers to an event in which possession of one object is superceded by possession of another. Then if we analyze the concept of 'possession', we find that it is similar or identical to the relationship expressed by the verb *has*, which it appears cannot be represented by a set-theoretic feature model. It follows that there can be no set of independent defining features for the category *money*; rather, membership in this, and many other categories (e.g., *toy, government, pet, game*) is defined by a relational decision rule. Such rules cannot be represented in terms of sets of independent features. Consequently, it appears that set-theoretic feature models, such as the feature-comparison model, are not sufficiently powerful to account even for category membership.

It might be argued that any theory of sense relations will ultimately be reducible to a theory of reference. Indeed, the initial plausibility of the feature-comparison model is largely dependent on the fact that it has been mainly applied to categories with clear perceptual referents, so that the postulated features have referential meaning. For example, one might suppose that many of the defining features of *canary* can literally be seen (e.g., 'yellow', 'small', 'flies', 'sings', etc.). However, Smith et al., have not answered the classic objections to the assumption that sense relations can be explained by a theory of reference (Frege, 1952; Katz, 1972). Accordingly, we presume that these objections stand against the feature-comparison model if it is considered to be such a theory.

It might also be asked, of course, whether network models can be extended to represent semantic relations other than containment (e.g., possession). While this is an open question, there is reason to believe they may. For instance, several network theorists (Anderson & Bower, 1973; Rumelhart, Lindsay & Norman, 1972; Schank, 1972) have used labeled relational arrows in order to represent different formal relationships between components. In linguistics, Katz (1972) has presented detailed proposals for representing the meanings of complex concepts in forms equivalent to graph structures. Thus while no network theory has yet been developed to deal with the full complexity of meaning, it at least may have the potential. Given the theoretical deficiencies of current set-theoretic models, it seems likely that if such models are elaborated they will become equivalent to a network representation.

But at the same time, it is clear that network models also face numerous serious difficulties in accounting for meaning, some of which have been mentioned in passing. The central issue is how network representations can be constrained. The distinction between set-theoretic and network models is, at heart, a difference in the conceptual resources that the two allow. Set-theoretic models are conceptually meagre and bare-boned, and as a result seem unable to capture the variety of tricks the human memory seems capable of performing. Network models face quite a different problem, for there are far fewer constraints on the kinds of representations they allow. The marker-search model has attempted to establish a few empirical constraints for a tiny fragment of English. While this attempt has met with some success, problems remain even here. If any such model is to be extended to other areas of language, it is essential that additional constraints be found.

This paper has focussed directly on the relative merits of two particular classes of psychological models of semantic memory. Nevertheless, we believe this discussion has significance for linguistics and philosophy as well, since it provides evidence that is inconsistent with some theories in these domains while demonstrating an empirical basis for others. If no psychological process model that can account for sentence verification results is consistent with the claim that truth is a continuous dimension, then a linguistic competence model based on that claim is unlikely to be justified. On the other hand, the psychological evidence that supports the marker-search model provides a broader empirical basis for the existence of those constructs (markers) that Katz (1972) needs to construct his definition of analyticity. The choice between models thus reflects not only narrow concerns within psychology, but also basic disagreements about the nature of human understanding.

REFERENCES

Anderson, J. R., & Bower, G. H. (1973) Human Associative Memory. Washington, V. H. Winston & Sons.

Bierwisch, M. (1969) On certain problems of semantic representation. *Found. Lang., 5,* 153–184.

Clark, H. H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. verb. Learn. verb. Beh., 12,* 335–359.

Clark, H. H., & Chase, W. G. (1972) On the process of comparing sentences against pictures. *Cog. Psychol.*, *3*, 472–517.
Collins, A. M., & Quillian, M. R. (1969) Retrieval time from semantic memory. *J. verb. Learn. verb. Beh.*, *8*, 240–248.
Frege, G. (1952) On sense and reference. In Geach, P., & Black, M. (Eds.), *Translations from the Philosophical Writings of Gottlob Frege.* Oxford, Basil Blackwell & Mott.
Glass, A. L., & Holyoak, K. J. (1974) The effect of *some* and *all* on reaction time for semantic decisions. *Mem. Cog.*, *2*, 436–440.
Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974) Production frequency and the verification of quantified statements. *J. verb. Learn. verb. Beh.*, *13*, 237–254.
Holyoak, K. J., & Glass, A. L. (In press) The role of contradictions and counter-examples in the rejection of false sentences. *J. verb. Learn. verb. Beh.*
Katz, J. J. (1972) *Semantic Theory.* New York, Harper & Row.
Kintsch, W. (1972) Notes on the structure of semantic memory. In Tulving, E., & Donaldson, W. (Eds.), *Organization of Memory.* New York, Academic Press.
Lakoff, G. (1972) Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Papers from the Eighth Regional Meeting, Chicago Linguistics Society.* Chicago, University of Chicago Linguistics Department.
Landauer, T. K., & Meyer, D. E. (1972) Category size and semantic memory retrieval. *J. verb. Learn. verb. Beh.*, *11*, 539–549.
Loftus, E. F. (1973) Category dominance, instance dominance, and categorization time. *J. exp. Psychol.*, *97*, 70–94.
Meyer, D. E. (1970) On the representation and retrieval of stored semantic information. *Cog. Psychol.*, *1*, 242–300.
Meyer, D. E., & Schvaneveldt, R. W. (1971) Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *J. exp. Psychol.*, *90*, 227–234.
Minsky, M., & Papert, S. (1969) *Perceptrons.* Cambridge, The M.I.T. Press.
Posner, M. I. (1970) On the relationship between letter names and superordinate categories. *Q. J. exp. Psychol.*, *22*, 279–287.
Rabbitt, P. M. A. (1966) Errors and error correction in choice response tasks. *J. exp. Psychol.*, *71*, 264–272.
Rips, L. J., Shoben, E. J., & Smith, E. E. (1973) Semantic distance and the verification of semantic relations. *J. verb. Learn. verb. Beh.*, *12*, 1–20.
Rosch, E. R. (1973) On the internal structure of perceptual and semantic categories. In T. M. Moore (Ed.), *Cognitive Development and Acquisition of Language.* New York, Academic Press.
Rumelhart, D. E., Lindsay, P. H., & Norman, D. A. (1972) A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization and Memory.* New York, Academic Press.
Schaeffer, B., & Wallace, R. (1970) The comparison of word meanings. *J. exp. Psychol.*, *86*, 144–152.
Schank, R. C. (1972) Conceptual dependency: a theory of natural language understanding. *Cog. Psychol.*, *3*, 552–631.
Smith, E. E. (1967) Effects of familiarity on stimulus recognition and categorization. *J. exp. Psychol.*, *74*, 324–332.
Smith, E. E., Rips, L. J., & Shoben, E. J. (In press) Semantic memory and psychological semantics. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 8. New York, Academic Press.
Smith, E. E., Shoben, E. J., & Rips, L. J. (1974) Structure and process in semantic memory: a featural model for semantic decisions. *Psychol. Rev.*, *81*, 214–241.
Sternberg, S. (1969) The discovery of processing stages: extensions of Donders' method. In W. G. Koster (Ed.), *Attention and Performance II.* Amsterdam, North-Holland Publishing Company.
Wilkins, A. T. (1971) Conjoint frequency, category size, and categorization time. *J. verb. Learn. verb. Beh.*, *10*, 382–385.

Résumé

On discute le fait que les théories de la mémoire sémantique suivent une divergence parallèle à celle des controverses concernant la représentation de la signification. Le modèle de comparaison des traits (Smith, Shoben et Rips, 1974) applique la théorie linguistique de Lakoff (1972) pour prédire le temps de réaction dans la vérification des phrases, alors que le modèle de recherche des marques, décrit ici, utilise le type de représentation sémantique défini par Katz (1972) pour expliquer des donnés analogues. Les deux modèles sont décrits et leur portée est revue. Le modèle de recherche de marques se vérifie bien mais en revanche une prédiction majeure de modèle de comparaison des traits est infirmée.

On discute le fait que le modèle de comparaison des traits est inadéquat pour rendre compte de la représentation sémantique tant que sa conception des consituants sémantiques reste inchangée.

Alternative conceptions of semantic theory*

ARNOLD L. GLASS
KEITH J. HOLYOAK

Stanford University

Abstract

It is argued that theories of semantic memory have diverged in a manner that parallels current linguistic controversy concerning the representation of meaning. The feature-comparison model (Smith, Shoben & Rips, 1974) applies the linguistic theory of Lakoff (1972) to predict people's reaction times to verify sentences, while the marker-search model, described here, uses the type of semantic representation outlined by Katz (1972) to explain a similar range of data. The two models are described and the evidence for each is reviewed. Available evidence supports the marker-search model, but disconfirms a major prediction of the feature-comparison model. It is argued that the feature-comparison model is in principle inadequate as a model of semantic representation, unless its conception of semantic components is substantially altered.

Philosophers and linguists have long discussed how the meaning of a word is represented in memory. In psychology, semantic memory research has approached this question by investigating possible mechanisms by which people use their knowledge about words to determine whether sentences are true or false. The dependent variable of major interest has been reaction time (RT) to verify simple propositions, such as *A dog is an animal.*

In a recent review paper, Smith, Shoben, and Rips (1974) have proposed a model, called the *feature-comparison* model, to account for the bulk of the RT differences reported in the semantic memory literature. This model emerges from what they term a 'set-theoretic' tradition in semantic memory research (Meyer, 1970; Schaeffer & Wallace, 1970). It uses a semantic theory outlined by Lakoff (1972) as the basis for a psychological process model. Set-theoretic models of semantic memory have been contrasted with net-

*The ordering of authors is haphazard. This paper has benefited from the extensive suggestions of Gordon H. Bower, Daniel Osherson and an anonymous reviewer. We are especially grateful to our good friends Edward E. Smith, Edward J. Shoben, and Lance J. Rips for the free exchange of data and ideas upon which this paper depended.

This paper was completed while A. Glass held an N.S.F. graduate fellowship and K. Holyoak held a Stanford University fellowship, and it was supported by Grant MH13950-06 from the National Institute of Mental Health to Gordon H. Bower.

work models, which represent the meaning of a word as a mapping between the word and a relational network (Collins & Quillian, 1969). However, the distinction between set-theoretic and network models has never been clearly drawn. Smith et al., describe set-theoretic models as those models in which concepts are represented by sets of elements. But network models can also define concepts componentially, so that at this general level the notations of set theory and of graph structures are largely interchangeable. In order to contrast the two types of models more clearly, we will describe two assumptions that can be used to distinguish set-theoretic from network models.

First, set-theoretic models restrict themselves to a very simple formal representation. Each element in the set representing a concept is treated as an atomic unit. Such formal devices as redundancy rules (Katz, 1972), which would permit one element to dominate (and therefore entail) another are excluded from the representation. Semantic relations are defined in terms of operations such as set inclusion; e.g., it might be assumed that a person can verify that a *dog* is an *animal* by determining that the set of features defining *dog* contains the set of features defining *animal*. In contrast, network models can include the graph-theoretic equivalents of redundancy rules in order to mark entailments, while the equivalents of antonymous n-tuples (Katz, 1972) can be used to mark contradictions between semantic components.

A second assumption, one that is central to the feature-comparison model, is that the relation of category membership is in some sense a matter of *degree*. This assumption is identical to Lakoff's (1972) hypothesis that absolute notions of truth and falsity should be replaced by a continuous truth dimension. This view would represent a sentence such as *A bat is a bird* as having some 'intermediate' truth value. In contrast, semantic relations in a network model are basically all-or-none: A component either dominates another, or it does not; and a component either contradicts another, or it does not. This type of representation therefore naturally leads to absolute rather than continuous notions of truth and falsity, as is advocated by Katz (1972). Under this view, a person might be uncertain as to the truth value of *A bat is a bird*, due to his ignorance or the sentence's ambiguity; but nevertheless, an absolute dichotomy remains between truth and falsity. Clearly, these parallel debates in psychology and linguistics are related to the extent that the goals of the two fields converge. Accordingly, the issues which distinguish these two classes of models have implications for linguistics as well as for psychology.

The intent of the present paper is to analyze the empirical and theoretical justification for these two types of models. The first section of the paper provides a critical review of the major evidence for the feature-comparison model of Smith et al. We are focusing our attack on the feature-comparison model for several reasons. First, it formulates the set-theoretic assumptions that we wish to call into question more clearly than any previous proposal. Second, the model has been given a precise formulation that allows the possibility of disconfirmation. Third, as Smith et al., point out, the feature-comparison model has been more successful in accounting for available verification data than any other set-theoretic model yet proposed. Accordingly, if we can show the Smith et al., model to be inadequate we will be rejecting not simply an arbitrary version of a set-theoretic model, but the most successful one yet devised.

The feature-comparison model invokes two separate mechanisms for verifying a sentence. In Section 2, by contrast, we present a network model that adapts the theory of Katz (1972) to psychological prediction, and which assumes a single basic mechanism for verifying sentences. In this alternative model the single underlying variable that determines both true and false RT is the time required for the person to access information that logically confirms or contradicts the truth of the presented sentences. Section 3 presents data that provides support for our proposal, while disconfirming a critical prediction of the feature-comparison model. The final section of the paper then examines broader theoretical issues concerning the nature of semantic representation that are raised by a comparison of the two models. We shall argue that a set-theoretic representation is in principle inadequate as a model of semantic memory.

1. The feature-comparison model: Review and critique

The feature-comparison model assumes that the meaning of a word is represented by a set of features, and that "some features will be more defining or essential aspects of a word's meaning, while others will be more accidental or characteristic features" (Smith et al., 1974, p. 4). Each feature is thus stored along with a weight indicating its degree of 'definingness' for the concept in question. The feature-comparison model posits two distinct serial stages that are used to verify sentences of the form *An S is a P*. In the first stage, the overall relatedness of the subject and predicate words is assessed in terms of all features (regardless of their definingness weights) of the two categories. If the overall relatedness of the subject and predicate words exceeds an upper criterion, a quick 'true' response is made. If their relatedness falls below a lower criterion, a quick 'false' response is made. Only if the overall relatedness falls between the upper and lower bounds is the second stage executed, resulting in a longer RT. This second stage separates the more defining features from the characteristic ones on the basis of feature weights, and compares only the more defining features of the subject and predicate. A 'true' decision is made in case all the defining features of the predicate are contained in the subject; otherwise, the decision is to respond 'false'. The feature-comparison model does not specify any relationship between overall semantic relatedness and the duration of second-stage processing. This model predicts that for true sentences, as the relatedness between subject and predicate increases, the percentage of quick stage-one 'true' responses will increase, resulting in faster mean RT for more related true sentences. But for false sentences, high relatedness will decrease the percentage of quick stage-one 'false' responses, resulting in slower mean false RT as relatedness increases.

We will consider whether the feature-comparison model's assumption of two serial processing stages is justified. This may be done in light of a test derived from one proposed for stage models by Sternberg (1969); viz., are there two conceptually distinct variables, one derived from stage one and the other derived from stage two, that both affect RT as predicted by the feature-comparison model? Smith et al., identify two variables that might satisfy this criterion for the feature-comparison model: (1) Semantic relatedness, which should affect the outcome of stage one only; and (2) category 'size' (i.e., the number of features that define a particular category), which should affect the duration of stage two only. Let us examine the evidence concerning these variables.

Semantic relatedness

For true sentences the feature-comparison model predicts that high semantic relatedness will result in a greater probability of a correct stage-one response, and hence, lead to relatively fast mean RT. Smith et al., review several studies showing that high relatedness indeed speeds up correct classification of an instance as a member of the test category. However, such evidence is open to an alternative interpretation. The problem is that rated relatedness has proved in every case to be positively correlated with the frequency with which the instance is produced as an association to the category name, as measured by association norms such as those of Battig and Montague (1969) (see Rips, Shoben, & Smith, 1973; Rosch, 1973; Smith, 1967; Smith et al., 1974; Wilkins, 1971). For instance the correlation between relatedness and production frequency (one standard measure of association strength) was 0.85 in the Rips et al., study. While this correlation is consistent with Smith et al.'s claim that production frequency reflects semantic relatedness, other evidence demonstrates that production frequency has an independent effect on RT. This evidence is provided by Experiment I of Smith et al., where the correlation between semantic relatedness and production frequency was only 0.49. In that study, production frequency was clearly a better predictor of RT than were relatedness judgments.

Smith et al., can argue, of course, that production frequency simply measures the underlying conceptual variable of relatedness more accurately than do ratings. However, other empirical results are difficult to reconcile with the notion that production frequency can be identified with relatedness. Loftus (1973) obtained measures of the production frequency (PF) of the category given the instance as a stimulus, as well as of the frequency of the instance given the category as a stimulus. She then varied whether the category or the instance was presented to the subject first in a verification task requiring determination of whether the instance was a member of the category. When the instance preceded the category (e.g., *robin-bird*), the instance-to-category PF determined RT; but when the category preceded the instance (e.g., *bird-robin*), the category-to-instance PF determined RT. In terms of the feature-comparison model, this result implies that for the same two words, relatedness differs depending on the presentation order. However, the model does not specify the exact composition rule that is to be used to compute overall similarity, and it is unclear whether or how such a rule can be made sensitive to word order. In other words, if production frequency is to be taken as a measure of relatedness, then the notion of relatedness will have to be considerably complicated.

Other conceptual problems arise in trying to identify production frequency with relatedness. When a person is asked to rate the relatedness of a pair of words, he is being asked to do what Smith et al., assume is done during stage one of sentence verification — i.e., to compare the subject and predicate words and assess their degree of relatedness. Production tasks, on the other hand, involve the retrieval of one concept given another as a cue. While Smith et al., may assume that these tasks measure relatedness, there is no *a priori* reason to believe this to be true. In section 3 we will argue that production frequency reflects a different conceptual variable, namely, the order in which information about word meanings is retrieved. Note that the results of Loftus, described above, have a straightforward interpretation in terms of retrieval. Suppose that verification in the Loftus paradigm requires that the person find a path between the two presented concepts (the category and the instance), beginning at whichever concept is presented first. Then the obtained effects of presentation order simply indicate that the frequency with which the second word is generated as a response to the first in a production task measures how quickly a path can be found from the first concept to the second during verification. Other evidence that production frequency is best conceptualized as a measure of the order of information retrieval is reviewed in Section 3.

Other experimental evidence, reviewed by Smith et al., generally supports their prediction that high relatedness increases false RT. Several studies have found that the RT to reject meaningful (high-related) false sentences (e.g., *All grains are wheats*) is longer than the RT to reject relatively anomalous sentences (e.g., *All typhoons are wheats*) (Kintsch, 1972; Meyer, 1970; Rips et al., 1973; Wilkins, 1971). However, the issue is not yet closed. A study by Glass, Holyoak and O'Dell (1974) suggests that false RT is not monotonically related to overall relatedness of the subject and predicate terms. Contrary to the Smith et al., prediction, false sentences in which the subject and predicate were very closely related (e.g., *Many arrows are dull*) were rejected *more* quickly than relatively meaningful sentences in which the subject and predicate were less related (e.g., *Many arrows are wide*). However, minimally-related anomalous sentences (e.g., *Many arrows are intelligent*) were rejected most rapidly of all. This latter result is not inconsistent with the earlier findings, since previous studies compared a mixture of relatively meaningful sentences, differing in relatedness, to anomalous (very low-related) sentences. Further false RT data that is incompatible with the feature-comparison model is reviewed in section 3 below.

Category size

Smith et al., also specified a variable which supposedly affected only stage two in their model, namely, category size. Since larger, more abstract categories logically have fewer defining features, fewer comparisons should be required in stage two in order to match defining features of a large category with the features of the test instance. Accordingly, holding constant the probability that stage two processing occurs (by controlling semantic relatedness), an increase in category size should decrease decision time. This prediction has been tested by studies that have varied category size while attempting to hold relatedness constant (Landauer & Meyer, 1972; Wilkins, 1971). But contrary to the prediction of the feature-comparison model, both studies found that statements involving larger categories took *longer* to verify than statements involving smaller categories (the difference was 32 msec for false sentences and 17 msec for true sentences, respectively). Smith et al., criticize these studies, arguing that neither study reported tests of the obtained differences against item variability (Clark, 1973). It has not been demonstrated, therefore, that this category-size effect is reliable. However, the fact that these trends are opposite to the prediction of Smith et al., remains problematic.

Finally, Experiment I in Smith et al. tests their category size prediction while escaping methodological problems inherent in the earlier studies. That experiment varied category size and semantic relatedness (as measured by production frequency) independently. The true RT results showed that production frequency was a highly significant variable; but when the variance in RTs attributable to production frequency was eliminated (by an analysis of covariance) the residual effect of category size did not approach statistical significance. Apparently, the conclusion best supported by available results is that, contrary to the prediction of the feature-comparison model, category size in itself has no effect on true RT when production frequency is controlled. Since category size is the sole variable so far proposed to affect stage-two processing, one must conclude that the two-stage model does not meet the evidential standards for multi-stage models proposed by Sternberg (1969).

Smith et al., also test the feature-comparison model by fitting a mathematical model of its major assumptions to the data of their second experiment. The mathematical model provided estimates of such parameters as the length of the duration of stage two in relation to category size, and the subject's criterion (based on relatedness) for making a response without stage two processing. They used two different estimation procedures, one based on relatedness ratings (with sixteen parameters) and one based on error rates (with ten parameters). When RT was predicted from the model on the basis of semantic relatedness ratings the obtained fit was extremely poor, with a correlation between predicted and obtained RT of only $r(14) = 0.69$. Furthermore, the fit provided by the more successful procedure seems to rest on the use of a general correlation between higher error rates and slower RTs. We tested this possibility by predicting RTs for Smith et al.'s data directly from the observed error rates. For this purpose we grouped both the 96 true items and the 96 false items into ten levels of error rates, so as to have at least five items at each level, and used linear regression to predict the mean RTs. For true items the correlation between predicted and observed RT was $r(8) = 0.972, p < 0.01$, and the root mean square deviation equalled 11.3 msec; while for false items the correlation was $r(8) = 0.929, p < 0.01$, with a mean deviation of 16.4 msec. Caution is necessary in

comparing these results with those of Smith et al., since we predicted two sets of ten mean RTs, while Smith et al., predicted a single set of 36 mean RTs. Nevertheless, such a comparison is suggestive. Smith et al., estimated ten parameters in order to predict RT, and obtained a correlation of $r(24) = 0.945$ between predicted and observed RT, with a mean deviation of 28.9 msec. Our predictions, each set of which is based on just two estimated parameters, are no less accurate than those obtained by Smith et al., using their more elaborate model and parameter-estimation procedure. Thus while it is true that the parameter estimates obtained by Smith et al., are consistent with the feature-comparison model*, their RT and error rate data are also consistent with the large class of models that predict a positive correlation between error rates and RT. While this correlation is indeed predicted by the feature-comparison model (see Smith et al., 1974), it is in fact a general empirical result commonly obtained not only in semantic-memory studies but in other RT studies as well (e.g., Clark & Chase, 1972; Meyer & Schvaneveldt, 1971; Posner, 1970). RT and error rate are generally taken to be convergent measures of item difficulty. Consequently, the burden of proof remains with Smith et al., to demonstrate that this relationship reflects processes specific to semantic decision-making, rather than more general response strategies typically used by subjects in RT experiments.**

It should also be noted that the data to which Smith et al., fit their model are drawn from an experiment of a rather problematic design. The subjects' task was to decide whether an instance was a member of a target category (e.g., *bird*), but all distractor

*For instance, Smith et al., found that the parameter estimate for stage-two duration was longer for small categories (280 msec) than for large ones (161 msec), as the feature-comparison model predicts. However, it is possible that this result was artifactual, since category size was confounded with category discriminability in their experiment. For the small categories of *fruit* and *vegetable*, subjects rated the vegetable instances used as more closely related to *fruit* than to *vegetable* (!), while they rated the *fruit* instances used as nearly as close to *vegetable* as to *fruit*. Clearly, *S*s had problems discriminating between what the *E*s called 'true' and 'false' instances for these categories; consequently, *S*s' RTs were slower for these 'small' categories than for statements about instances of the large categories, *animal* and *plant*. However, Smith et al., did not introduce a parameter to account for category discriminability. This difference between the categories therefore had to be reflected by some other parameter. The most likely candidate for this role is the estimate of stage-two duration, since different parameters were estimated for the stage-two duration of large and small categories. Accordingly, the increase in RT for small categories resulting from the difficult discrimination between 'true' and 'false' instances may have been reflected in the parameter estimates by a longer estimate of stage-two duration for small categories. (We thank L. Glass for this suggestion.)

**Smith et al., also found that error RTs are faster than correct RTs, as their model predicts. However, it is not clear whether this is a general effect, as we have found no consistent relationship between error and correct RTs in data of our own. Even if the effect is general, it may be accounted for by a plausible strategy of how response speed might be traded off against accuracy. Subjects may tend to gradually speed up their responses over trials until they make an error. The occurrence of an error may then make the subject momentarily more cautious, and hence slower and more accurate (Rabbitt, 1966). Then he may begin a gradual decrease in RT until the next error. Such a cyclic pattern of response times would result in faster RTs for errors than for correct responses. Furthermore, since difficult sentences require more processing time for a correct decision, and thus are more likely to produce an error if the subject tries to respond quickly, this strategy would also produce a positive correlation between error rate and correct RT across conditions.

instances were drawn from a single non-target category (e.g., *insect*). For this example, the subject could therefore logically decide to respond 'true' not only by verifying that a given instance (e.g., *canary*) was a *bird*, but also by verifying that it was *not* an *insect*. Conversely, he could respond 'false' either after he verified that the instance (e.g., *termite*) was not a *bird*, or after he verified that it was an *insect*. The obtained pattern of RTs suggests that subjects in fact used all these possible decision strategies. True RT was found to depend not only on the relatedness of the instance to the target category (e.g., *canary* to *bird*), as the feature-comparison model would predict, but also on the relatedness of the instance to the non-target category (e.g., *canary* to *insect*). Both of these variables influenced false RT as well. Smith et al., offer only a *post hoc* explanation of these unexpected effects, suggesting that some subjects used different strategies than others, or that subjects varied their strategies from trial to trial (see their Footnote 9). It is clear, however, that these results were not predicted by the feature-comparison model, nor reflected in the parameters of the mathematical model. Nor is it clear that the pattern of results to which Smith et al., fit their model would generalize to a situation in which distractor items were drawn from a variety of categories.

In summary, the feature-comparison model predicts that the time used to make semantic decisions will be determined by two variables: Semantic relatedness and category size. However, these predictions are not unambiguously supported by available data. Production frequency appears to predict true RT more accurately than rated relatedness, while category size does not predict RT at all.

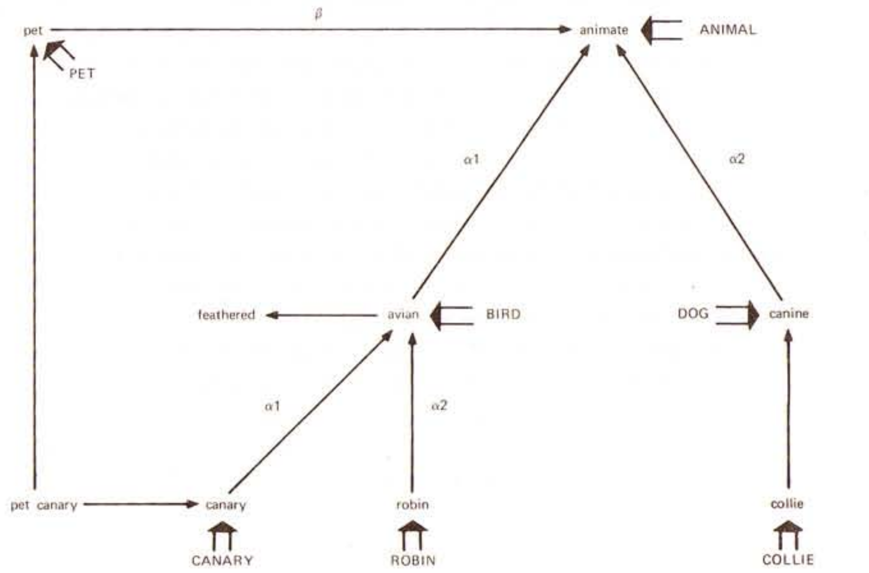
In addition to these empirical difficulties, there are a number of further conceptual problems with the feature-comparison model. A major problem is that the model does not appear to be specified in terms of explicit mechanisms. For example, what mechanisms might plausibly allow a holistic comparison of the type that is postulated to occur in stage one? How are subject and predicate features matched so that a hypothetical overall relatedness estimate can be calculated swiftly? This process must presumably take place without identifying feature dimensions, or else it would seem that a more reasonable and parsimonious strategy would be to compare defining features of the subject and predicate immediately, as is done in stage two of the hypothesized process. Until these questions have been answered, it is impossible to tell how the first stage of the model could ever be executed.

2. An alternative approach: Ordered marker-search

Semantic representation

We will now describe a marker structure (Katz, 1972) in sufficient detail to account for the verification of sentences quantified by *all* or *some* (e.g., *All canaries are birds*, *Some birds are canaries*). These types of sentences are logically equivalent to the categorization tasks to which the feature-comparison model has been explicitly applied (e.g., deciding that *canary* is an instance of the category *bird*, or that the category *bird* contains the instance *canary*). No attempt will be made to describe how relationships other than category membership could be represented, such as possession (*has*) or ability (*can*). Thus the only verb which can be represented in the structure to be described is the present tense of the copula (*is, are*). Discussion of the extendability of the two models will be postponed until the final section of this paper.

Figure 1. *Hypothetical marker structure, illustrating word-to-marker and marker-to-marker associations.*



A portion of such a marker structure is diagrammed in Figure 1. In this figure labels representing words are in capital letters, while labels of markers are in small letters. Thus BIRD represents a word, while (avian) stands for a marker. Where possible, markers are labeled with adjectives to emphasize that they are best thought of as properties (e.g., (animate)), rather than as categories or exemplars. For many markers no appropriate English adjective with which to label them exists; these are labelled with nouns. Thus the marker labeled (canary) stands for an abstract concept roughly equivalent to "possessing the essential properties of a canary".

As Figure 1 illustrates, we assume that most common words are directly associated with a single marker in the attribute structure. In each case we will refer to this marker as the 'defining' marker for the particular word. For example, (avian) is the defining marker for *bird*. In the case of words with multiple meanings (e.g., *bank*), each sense of the word will access a different defining marker. Note that no single word is associated with the marker labeled (pet canary). This illustrates our assumption that only a subset of the markers in memory are directly associated with words. But we assume that all markers can potentially be accessed during the search procedure used to verify sentences.

The marker-search theory has two basic structural assumptions. First, we assume that markers are interrelated in such a way that any marker stands for, or dominates, a set of further markers associated with it. In Figure 1 these associations are represented by arrows. For example, the arrows pointing from (avian) to (animate) and (feathered) indicate that (avian) stands for the set {(animate), (feathered)}. The property of containment represented by the arrows is transitive. Thus in Figure 1, since the marker (robin) implies – that is, dominates – the marker (avian), it must also dominate (animate) and (feathered). In other words, (robin) stands for the set of markers {(avian), (animate), (feathered)}. Also, by definition any marker dominates itself.

The arrows in Figure 1 are labeled in order to illustrate the second basic structural assumption of the model – that information about contradictions is represented in memory through the associations between markers. In the network structure a contradiction arises whenever two arrows labeled with the same greek letter meet at a common node. In the figure arrows which connect animal species to (animate) are labeled with an α, while the arrow connecting (pet) to (animate) is labeled with a β. Thus a contradiction arises at (animate) between (avian) (α_1) and (canine) (α_2), but not between (avian) and (pet) (β). This notational system simply represents our intuitive knowledge that a bird cannot be a dog, but might possibly be a pet. Note that no significance is attached to the labels used (α, β, etc.). The essential point is simply that intersections are of two types – contradictory and non-contradictory. It should also be noted that in Figure 1 the marker (robin) dominates (avian) while (collie) dominates (canine). It therefore follows from the transitive properties of the arrows that a contradiction of the sentence *A canary is a collie* also arises at (animate).

This treatment of redundancy rules and antonymy is based on the proposals of Katz (1972). The present notation is a variant of that used by Katz, who introduces superscripted markers to form antonymous n-tuples. The notation we are using is similar to that suggested by Bierwisch (1969), and has the advantage of capturing the close relationship between redundancy rules and contradictions.

The different logical relationships between concepts that are denoted by the quantifiers *all* and *some* (*All S are P* and *Some S are P*) can be specified in terms of the relationships between the defining markers of the subject and predicate (see Table 1). An *All*-statement is true if the defining subject marker dominates the defining predicate marker. For example, since in Figure 1 an arrow points from (avian) to (animate) it follows that *All birds are animals* is true. Similarly, since (avian) also implies (feathered), *All birds are feathered* is true as well. The truth conditions for *some* are slightly different. A *Some*-statement is true if there exists a marker *from* which arrows lead to both the defining subject marker and the defining predicate marker. The sentence *Some animals are feathered* is therefore true, since (avian) satisfies the above criterion.

Table 1. *Decision criteria for true and false sentences quantified by* All *or* Some

	True sentences		False sentences	
Quantifier	Criterion	Examples	Criterion	Examples
All	Defining subject marker dominates defining predicate marker	All birds are animals.	Defining subject marker contradicts defining predicate marker	*Contradictory:* All birds are dogs.
			A marker which dominates defining subject marker contradicts defining predicate marker	*Counterexample:* All birds are canaries.
Some	A marker dominates both defining subject marker and defining predicate marker	Some birds are canaries. Some birds are pets.	Defining subject marker contradicts defining predicate marker	*Contradictory:* Some birds are dogs.

As discussed above, falsity based on contradiction can also be explicitly defined in terms of the relations among arrows. Both *All-* and *Some*-statements are false if a contradiction exists between the defining subject marker and the defining predicate marker; i.e., if similarly-labeled arrows from the two defining markers meet at a common node. For example, we have already seen that (avian) contradicts (canine); consequently, *All/Some birds are dogs* is false.

In addition, an *All*-statement can be falsified by a slightly different type of contradiction. An *All*-statement is also false if there exists a marker which dominates the defining subject marker, but contradicts the defining predicate marker. *All birds are canaries* is therefore false because (robin) satisfies the above criterion, and thus serves as a counterexample to the sentence – it stands for a bird that is not a canary. Notice that this type of false *All*-statement would be true if the quantifier were *some* (e.g., *Some birds are canaries*). This kind of false *All*-statement will be referred to as a "Counterexample" sentence. The distinction between Counterexample sentences and those falsified by a direct contradiction (Contradictory sentences) will play a critical role in predicting the time taken to reject false sentences.

An interesting property of this representation is that it allows the person to be uncertain about the truth of certain sentences. For example, the sparse network of Figure 1 does not contain the information that some dogs are pets, since there is no marker equivalent to (pet dog) which dominates both (dog) and (pet). On the other hand, (dog) and (pet) have a non-contradictory intersection at (animate), indicating that a dog *might* be a pet. So if this memory were probed with the sentence *Some dogs are pets*, neither the criteria for a 'true' nor for a 'false' decision could be satisfied, and the appropriate response would be something like "It's possible, but I'm not sure". A real subject might give a similar answer to a sentence about some really obscure fact such as *Some anteaters are pets*.

We should point out that our discussion of structural representation has left unresolved a number of serious problems associated with marker theory. We have nothing to say about what elements in the marker system are primitive, except to agree with Katz (1972) that this question cannot be decided *a priori*, but only after considerable empirical study. Other important issues that have been ignored concern how a marker system could be acquired developmentally. For instance, how are redundancy relations generated between markers, and what determines whether an intersection between two relations will be marked as contradictory? While the evidence we will discuss below is solely from studies of sentence verification by adults, the adequacy of any representational proposal cannot be firmly established without studies of how a system of information can be acquired.

Searching the marker structure

According to Katz (1972), the set of markers that form the dictionary entry for a word is unordered. Accordingly, while the hierarchical structure in Figure 1 serves to represent the logical relations between markers, it is not necessary to assume that it indicates the order in which the markers are accessed to verify a sentence. If this assumption *could* be made, it would provide a strong empirical and formal constraint on network representations (Collins & Quillian, 1969). However, there is now ample evidence that people sometimes access relatively abstract markers more quickly than those that are less abstract. For instance, people can decide that scotch is a drink more quickly than they can decide that it is a liquor (Smith et al., 1974). Examples such as these demonstrate that the order in which markers are accessed is not always hierarchical.

However, the fact that the marker set is unordered in Katz's theory does not mean that the markers are accessed in some random fashion. It is possible to specify a performance model based on ordered, though non-hierarchical, search procedures. This possibility can be realized by modifying the structural representation in Figure 1 to include additional redundancy-rule pathways. For example, we will allow the possibility of a direct link between (canary) and (animate), as well as the illustrated pathway from (canary) to (avian) to (animate). To establish an ordering between any two alternative pathways, we can assume either that one of the pathways has a higher probability of being searched first, or that both pathways are searched in parallel with one requiring less time to traverse.

Postulating additional redundancy rules immediately raises a serious question: How is the model to be constrained? Abandoning a strictly hierarchical representation leaves the model without any *formal* constraint that would prevent each node from being directly connected to every node that it dominates. It would be possible to fit the pattern of RTs from any sentence verification task *post hoc*, simply by adding to the structural representation whatever additional connections are required. The price of this freedom, of course, is that the model would be rendered entirely vacuous.

But while no *formal* constraints on the representation prevents this outcome, it may be possible to find *empirical* constraints that will yield specific, testable predictions. This is the strategy that is followed in the studies to be reviewed below. First an empirical measure of search order, independent of RT, is identified. Since the same representation is presumably used in all tasks that depend on the kind of semantic information we have described, the results from one such task should predict performance on the other. Secondly, verification of different sentences will sometimes depend on the *same* semantic information. The relative speed with which a particular bit of information can be accessed should therefore determine the RT to verify a number of different sentences. In general, the complexity of the representation that will be required to account for sentence verification is an empirical issue. If it eventually becomes necessary to postulate unlimited redundant connections in the network, the marker-search model will be rejected as unworkable. But if sufficiently strong empirical constraints can be maintained, the model will have explanatory value.

The marker-search model takes the order in which markers are searched to be the underlying variable determining differences in semantic decision time. In the studies discussed below, we assume that the search procedure is subject to the following constraints: (1) During sentence verification the markers accessed by a word include not only those markers dominated by the defining marker of the word, but also those which dominate the defining marker. For example, consider the word *bird* in Figure 1, which is defined by the marker (avian). We assume that the marker search-set for the word *bird* will include both (animate) (which is dominated by (avian)), and (canary) (which dominates (avian)). This assumption is necessary to explain how the various types of true and false sentences summarized in Table 1 can be verified. Decisions about true *All*-statements and Contradictory false sentences are based on a marker dominated by the defining subject marker; decisions about true *Some*-statements and Counterexample *All*-statements, on the other hand, are based on a marker dominating the defining subject marker. (2) The order in which the markers accessed by a word are searched is independent of the particular quantifier or the truth value of the sentence in which the word appears. Thus the markers of *building* will be accessed in the same order in sentences such as *Some buildings are houses* and *All buildings are houses*. (3) The search will self-terminate as soon as either the criterion for a 'true' or for a 'false' response (as discussed above) is satisfied. For example, a 'true' response will be made to the sentence *All birds are animals* as soon as (avian) is found to dominate (animate); while a 'false' response will be made to the sentence *All birds are dogs* as soon as (avian) is found to contradict (canine).

It would presumably be possible to construct a number of different explicit search mechanisms consistent with the above constraints. These might differ in features such as the degree to which search is serial or parallel. However, the predictions outlined below will not discriminate between the various possible search models based on the semantic representation and search constraints described.

3. Evidence for the marker-search model

In this section we review the evidence for the search model provided by recent experimental results. The model takes the order in which markers are searched to be the variable that determines differences in semantic decision time. In order to test this model, it was necessary to develop a measure, independent of RT, of the order of marker search. Such a measure was proposed by Glass et al. (1974). They asked subjects to provide true one-word completions for incomplete sentences of the form *All/Some S are*________, and tabulated the frequency with which different words were given as predicates. This constrained association technique is similar to the way production frequency norms are collected from subjects who are asked to produce different instances as responses to a category name (Battig & Montague, 1969). Glass et al., assumed that the frequency with which a word appeared as a completion reflected the probability with which its corresponding defining marker was accessed from the defining subject marker.

Clearly production frequency can be at best an imperfect measure of search order, since we have assumed that many markers will not correspond to single common English words. This problem is particularly acute in the case of anomalous false sentences, in which the subject and predicate words generally differ at the level of abstract markers such as (living) versus (non-living) (e.g., *All birds are chairs*). These sentences are typically rejected relatively quickly (Kintsch, 1972; Wilkins, 1971); but since the abstract markers on the basis of which they could be rejected seldom define common English words, production frequency is not a valid measure of the speed with which such markers are accessed. The problem of predicting RT to reject anomalous sentences is discussed more fully in Holyoak and Glass (1975). In the present paper we will only discuss sentences for which production-frequency measures make clear RT predictions; i.e., true sentences, Counterexample sentences, and Contradictory sentences in which the subject and predicate words differ with respect to relatively specific markers (e.g., *All birds are reptiles*).

The marker-search model predicts that true sentences with high production frequency (PF) will be verified more quickly than sentences with lower PF. Glass et al. tested this prediction for sentences with five different quantifiers (*All, Many, Some, Few* and *No*), and both noun and modifier predicates (e.g., *All birds are animals, All birds are winged*). In each case the corresponding PF norms successfully predicted RT. These results extended the findings of Loftus (1973) and Wilkins (1971), who also found that high PF leads to fast true verification RT. The Glass et al., findings have since been replicated by Glass and Holyoak (1974) and Holyoak and Glass (1975). We have also seen that in studies where semantic relatedness was a successful predictor of true RT, it was confounded with PF (Rips et al., 1972; Rosch, 1973). Furthermore, in the one case in which the effects of the two variables have been compared, PF was a much better predictor than was relatedness (Smith et al., 1974). Accordingly, the available data concerning true RT are consistent with the model.

Generation of false sentence completions

While previous experiments were primarily concerned with using PF (or relatedness) to predict true RT, the results obtained by Glass et al., for false *Many*-statements (discussed earlier) indicated that it should also be possible to use such norms to predict RT to reject false sentences. These predictions, to be outlined below, were tested by Holyoak and Glass (1975).

As a necessary initial step, Holyoak and Glass collected false PF norms for sentences quantified by *all* or *some*. They had 32 Stanford undergraduates generate *false* completions for sentences of the form *All S are*________ and *Some S are*________. The resulting false PF norms were compared with true PF norms compiled in previous work (Glass & Holyoak, 1974; Glass et al., 1974). Several striking relationships between true and false sentence completions provided evidence for the marker-search model. Referring to Figure 1, let us consider strategies that people might use to generate false completions of a sentence. One plausible strategy would be to first access a marker dominated by the defining subject marker, and then use it to compute a contradiction as a response. For example, when presented with *All/Some birds are______*, the person might follow the arrows from (avian) to (animate), and then from (animate) to (canine). This procedure would generate the false completion *dog*, producing a Contradictory sentence, as defined in Table 1. If this strategy were actually used by subjects, and search order is independent of the quantifier (as the model assumes), there should be a close relationship between the frequency with which *All/Some birds are dogs* is produced as a false sentence, and the frequency with which *All birds are animals* is produced as a true sentence. Specifically, we expected that each high-PF true completion of *All S are________* from the norms of Glass and Holyoak (1974) and Glass et al (1974), would determine some high-PF Contradictory completion of both *All S Are________* and *Some S are________*. As the examples given in the top of Table 2 illustrate, this prediction was confirmed. Fourteen of the 16 highest frequency true *All*-statements (produced by between 35% and 78% of the respondents) corresponded to high-PF Contradictory sentences (produced by from 19% to 56% of respondents).

Referring again to Figure 1, let us illustrate a second strategy that subjects might use to generate false *All*-statements. This strategy depends on our assumption that markers that dominate the defining subject marker are also accessed by the subject word. The subject can therefore go directly from the defining subject marker to a marker that dominates it, and then use this marker to produce a false completion. For instance, given the fragment *All birds are__________*, the person can simply follow the arrow from the (avian) to the (canary) marker, and respond with the word *canary*. This procedure would produce Counterexample sentences, such as those shown in Table 2. Note that exactly the same procedure would produce a *true* sentence if the quantifier were *some* (e.g., *Some birds are canaries*). Since the model assumes that search order is independent of the quantifier, our second prediction was that each high-PF true *Some*-statement (from the earlier norms of Glass and Holyoak and Glass et al.) would correspond to a high-PF Counterexample *All*-statement. This prediction was confirmed. The bottom of Table 2 lists four examples of the 22 highest frequency true *Some*-statements from the earlier norms (given by from 22% to 89% of respondents). Each of these 22 true *Some*-statements (e.g., *Some flowers are roses*) corresponded to a high-frequency false Counterexample sentence (e.g., *All flowers are roses*), given by from 16% to 53% of respondents.

A comparison of true and false sentence completions therefore supported the assumption that contradictions can be found in memory. People appear to access the same markers in producing false as well as true sentence completions, except that they contradict the marker in generating false completions.

Table 2. *Examples of relationships between high-frequency true and false sentences*

True	False
Contradictory falses	
All birds are animals.	All/Some birds are dogs.
All chairs are furniture.	All/Some chairs are tables.
All women are humans/females.	All/Some women are males.
All diamonds are stones.	All/Some diamonds are emeralds.
Counterexample falses	
True	False
Some flowers are roses.	All flowers are roses.
Some prisoners are men.	All prisoners are men.
Some books are novels.	All books are novels.
Some teachers are professors.	All teachers are professors.

Rejection of false sentences

The central evidence that discriminates between the marker search and feature-comparison models concerns RT to reject false sentences (Holyoak & Glass, in preparation). The marker-search model predicts that disconfirmation of meaningful false sentences (i.e., Contradictory and Counterexample sentences) requires discovery of a contradiction. Consequently, the sooner the person can access a marker that brings out a contradiction between the subject and predicate, the quicker such sentences will be rejected. For both Contradictory and Counterexample sentences, the order in which contradictions are discovered should be predicted by our production frequency norms; however, the variable that determines false RT should be quite different for these two kinds of false sentences.

Contradictory sentences (e.g., *All/Some birds are dogs*) contain predicates that directly contradict the subject (See Table 1). For such a sentence, its production frequency was taken as an index of the speed of accessing the marker (e.g., (animate)) that produces a contradiction between the subject and predicate. For Contradictory sentences, then, false statements that have high frequency in the norms should be rejected more quickly than false statements given with low frequency in the norms. In agreement with this prediction, Holyoak and Glass found that high-PF Contradictory sentences quantified by *all* or *some* were rejected significantly more quickly than low-PF Contradictory sentences (1319 versus 1468 msec). This result extended the previous findings for false *Many*-statements obtained by Glass et al. (1974). Furthermore, Holyoak and Glass had subjects rate the relatedness of the subject and predicate words for each of their false sentences, and found that their high-PF sentences were rated as significantly more related than their low-PF sentences. Since the feature-comparison model of Smith et al., (1974) predicts that high-related false sentences will be rejected *slower* than less related false sentences, these RT results are opposite to the prediction of the feature-comparison model.

In Counterexample sentences (e.g., *All birds are robins*), the predicate does not directly contradict the subject. The marker-search model therefore predicts that in order to reject this type of sentence, the person must discover some marker (representing an exemplar) that dominates the defining subject marker (e.g., (canary)) and that contradicts the predicate. Accordingly, the RT should be fastest for those sentences for which a disconfirming counterexample was produced most frequently as a true *Some*-completion. Specifically, the RT to reject a Counterexample sentence such as *All birds are robins* should be faster the higher the frequency with which the most common counterexample (e.g., 'canary') was given as a true completion of the sentence *Some birds are*________. Since there is no direct contradiction between the subject and predicate for this type of sentence, the production frequency of the sentence itself (*bird* to *robin*) should have no appreciable effect upon the time to reject it.

To test this prediction, Holyoak and Glass selected Counterexample sentences in which the PF of the most common counterexample was varied orthogonally with the PF of the sentence itself. As predicted, sentences with high-PF counterexamples were rejected significantly more quickly than sentences with low-PF counterexamples (1397 versus 1506 msec), while the PF of the sentence itself had no significant effect on RT. Since PF and semantic relatedness were again positively correlated, the feature-comparison model is unable to account for these results.

To summarize, experimental evidence supporting the marker-search model comes from several sources. The model accounts for observed semantic relationships between true and false sentence completions. It predicts the strong correlation between production frequency and RT to verify true sentences. Most strikingly, the model successfully predicts the RT to reject meaningful false sentences (both Contradictory and Counterexample sentences). The same data for false sentences disconfirm a major prediction of the feature-comparison model, namely, that false sentences with subject and predicate words closely related in meaning are necessarily slow to be rejected.

4. Issues in semantic representation

In this section we examine some theoretical issues that can be highlighted by a comparison of the feature-comparison and marker-search models. These issues center on the distinctions between set-theoretic and network models outlined earlier, and particularly on the differing conceptions of the representation of word meaning which emerge from the two approaches.

Semantic redundancy rules and category size

The marker-search model postulates abstract markers representing the entire set of elementary components associated with the definition of a word (e.g., the marker (avian) would dominate all the markers that define the word *bird*). The markers that define the subject and predicate can then be matched directly during sentence verification; e.g., to verify *All canaries are birds* the system need only discover that a set inclusion relation holds between (canary) and (avian). But should it be necessary, more elementary components can also be recovered by means of associations representing semantic redundancy rules (Bierwisch, 1969; Katz, 1972). Thus, since a set-inclusion relation holds between (canary) and (avian), and also between (avian) and (animate), the sentence *All canaries are animals* can also be verified. While the latter sentence requires a longer search, the final match that allows a decision is again between just two markers ((canary) and (animate)). Accordingly, the 'size' of a category, in terms of the number of elementary components associated with the defining marker of a word, is not a relevant variable in predicting RT.

In contrast, the feature-comparison model does not postulate markers that represent sets of meaning components. The features of such a model are not structured hierarchically. This theoretical distinction yields a very different description of the verification process than that given by the marker-search model. In the second stage of the feature-comparison model, verification of *All canaries are birds* would require matching the set of defining features of *bird* (e.g., 'feathered', 'egg-laying', 'breathing', 'solid', etc.) with the defining features of *canary*. If the predicate were *animal*, its feature list would be shorter, so that second-stage processing should require less time. But as indicated in the first section, there is no evidence that this prediction holds. The introduction of redundancy rules as a psychological construct thus serves at least two functions: It simplifies the search process presumed to occur during verification, and it explains the negative finding that sentences with relatively abstract predicate categories are not verified any more quickly than sentences with less abstract predicates, as long as production frequency is controlled.

The defining/characteristic distinction

A related issue concerns Smith et al.'s proposal that features can be weighted by their degree of 'definingness', ranging from clearly defining to simply characteristic features. They cite Lakoff's (1972) analysis of hedges as major linguistic support for this defining versus characteristic distinction among features. In particular, three hedges are alleged to clearly differentiate the types of features shared by the subject and predicate of certain sentences. According to their analysis, the sentence *A robin is a true bird* is acceptable because *robin* shares both the defining and characteristic features of *bird*, since the features characteristic of a category are those that define common instances. The sentence *Technically speaking, a chicken is a bird* is acceptable because *chicken* shares the defining but not the characteristic features of *bird*. Finally, the sentence *Loosely speaking, a bat is a bird* is acceptable because *bat* shares the characteristic features of *bird*, but not the defining ones. If the hedge in any of these examples is replaced with one of the other hedges (e.g., *Technically speaking, a bat is a bird*), the resulting sentence is less acceptable, since the subject and predicate do not share the type of features specified by the hedge. Smith et al., conclude that this linguistic result is best explained by assuming that features differ in the degree to which they define words.

However, one may explain such hedges in another way. It seems that certain common English words have at least two definitions. One is a popular definition, often learned early in life, such as the fact that a bird is a small flying animal with wings. The other is a technical definition, first agreed upon for some specific purpose by scientists or lawyers, and eventually picked up by dictionary writers and imposed upon the general public. The biological definition of a bird is an example of a technical definition. When an instance fulfills the requirements of both the technical and popular definitions, it may be said to be 'a true' member of the category (as *robin* is for *bird*). When an instance fulfills only the technical definition (e.g., *chicken* for *bird*), then we say 'technically speaking'; when only the popular definition is satisfied (e.g., *bat* for *bird*), we say 'loosely speaking'.

This explanation of hedges can account for the unacceptability of certain sentences that create problems for Smith et al.'s defining/characteristic explanation. For instance, if we use a category as the subject of a sentence, and a 'true' instance as the predicate, the subject will still share characteristic features with the predicate. Accordingly, the hedge 'loosely speaking' should apply, producing such examples as *Loosely speaking, a bird is a robin*. But such sentences are unacceptable. A similar difficulty occurs with two 'true' instances of a category, which surely share many characteristic features. The Smith et al., analysis therefore predicts that a sentence such as *Loosely speaking, a robin is a canary* should be acceptable, but again it is not. The defining/characteristic explanation of hedges requires some additional assumption to explain these cases.* In contrast, the unacceptability of these latter sentences follows directly from our popular/technical explanation, since in neither of these cases does the predicate appear to have two definitions, nor does the subject satisfy any definition of the predicate.

Furthermore, one suspects that the as yet unspecified characteristic and defining features on which the hedges 'technically' and 'loosely' are supposedly based can never be specified. Do people know that the sentence *Technically speaking, a whale is a mammal* is true because a defining feature for *mammal* is shared by *whale*? Or do they know it is true simply because they know that there exists a technical definition of *mammal* that includes *whale*, even though they don't know what it is? In this case, the technical definition actually becomes definition by enumeration.

*One such additional assumption would be that the predicate category must always be more general than the subject category. However, it is not clear that an adequate metric of generality can be specified in terms of the feature-comparison model. One might suppose that more general categories are those with fewer defining features. But this tack allows comparisons of generality only when the categories are logically nested. But two instances, such as *canary* and *robin*, are not nested; consequently, the 'number of defining features' metric does not specify how we know that *canary* is not more general than *robin*. There is no apparent *a priori* reason to suppose that all instances have an equal number of defining features.

An explanation of these hedges in terms of popular and technical definitions has certain testable implications. For instance, we would not expect people in primitive 'non-technical' cultures to make a distinction equivalent to the difference between 'loosely speaking' and 'technically speaking'. Also, since children presumably acquire popular definitions of words earlier than technical definitions, young children should classify an instance as a 'correct' member of a category if and only if it satisfies the popular definition. For example, young children presumably would classify a bat as a bird, and a whale as a fish.

The status of network models

Smith et al., cite the evidence from hedges as a general source of difficulty for network models of semantic memory, of which the marker-search model is an example. They do point out that a network model could incorporate the notion of popular and technical definitions of words by including separate markers for the two definitions. But Smith et al., object to this tack as unparsimonious, presumably because it leads to a proliferation of markers. But it is not clear that this solution is less parsimonious than that offered by the feature-comparison model, which is to assume that a weight indicating 'definingness' is stored with every feature-category pair in memory. That is, given a model based on meaning components, is it more parsimonious to add a finite number of extra components (e.g., popular and technical definitions) or a completely new theoretical mechanism (e.g., definingness weights)?

We should make clear that every use of a hedge does not require a network model to postulate an additional definition for a word. Thus Smith et al., suggest that the acceptability of the sentence *A decoy is a fake bird* requires a network model to postulate an additional definition of *bird* representing 'pseudo-instances'. But in terms of the semantic theory of Katz (1972), the definition of *fake* would be represented by markers corresponding to 'intended to appear as'. Through syntactic and semantic amalgamation rules, a *fake bird* would thus be an object intended to appear as a bird. Any word with defining markers that contain the markers of this compound (such as *decoy*) would be correctly classified as a *fake bird*. This treatment of the meaning of a noun plus modifier could be incorporated into a network model. In contrast, Smith, Rips and Shoben (1974) claim that *fake* is used when the subject and predicate share only characteristic features. However, this proposal is clearly incorrect, as is demonstrated by the unacceptability of the sentence *A bat is a fake bird*. While a bat, just like a decoy, both looks like a bird and isn't a bird, its resemblance is essentially accidental rather than intentional. Consequently, this sentence is unacceptable. It is not clear, however, how the feature-comparison model could distinguish between the acceptability of these two examples simply on the basis of feature overlap.

A second criticism that Smith et al., direct at network models is the finding that typical instances are categorized more quickly than atypical ones (Rips et al., 1973;

Rosch, 1973; Smith et al., 1974). Smith et al., discuss two ways in which network models could incorporate this finding. It may be that intermediate nodes are interposed between the markers representing an atypical instance and the category (e.g., (chicken) might first imply (domestic bird), which then in turn implies (avian)). Also, a network model could account for the effect of typicality on RT simply by assuming that the defining marker of the category is higher on the search list for typical instances than for atypical instances.

Both these possibilities seem plausible. As Smith et al., point out, typicality effects pose difficulties only for those network models that assume that search order somehow mirrors the logical structure of the concepts (Collins & Quillian, 1969). But since the marker-search model does not make this assumption, it avoids these difficulties. To illustrate, consider how the marker-search model handles one typicality result obtained by Smith et al., which they view as particularly troublesome for network models. To use their example, they found that *robin* is more typical than *chicken* for the superordinate *bird*, whereas *chicken* is more typical than *robin* for the superordinate *animal*; moreover, this interaction is reflected in verification RT. This result can be represented very simply in terms of the marker-search model. The markers (chicken) and (robin) can both have direct associations to both (avian) and (animate). If search is serial, in each case there must be a trade-off in terms of whether the association to (avian) or to (animate) is searched first. Which ever has priority, the other must be accessed more slowly. Production frequency norms would presumably indicate that for (chicken) the association to (animate) is searched relatively early, whereas for (robin) the association to (avian) has priority. Consequently, people will be relatively quick to verify both *A chicken is an animal* and *A robin is a bird*, but slow to verify *A chicken is a bird* and *A robin is an animal*. Not only is this result consistent with the marker-search model, it in fact provides evidence for the assumption that search is reliably ordered.

The basic issue, therefore, is whether the effects of typicality are best explained by a model based on continuous variation in degree of category membership, or by a model based on the discrete nodes of a network structure. The linguistic and experimental results cited by Smith et al., do not discriminate between these two conceptions of semantic memory.

Can the feature-comparison model be extended?

In order to handle the RT results reviewed in the previous section, the feature-comparison model could incorporate ordered search strategies of the type we have described as an elaboration of the second processing stage. 'False' decisions could be based on the discovery of a contradiction, defined in terms of a relationship between features. But it would then appear that the first stage of the revised model would either be superfluous or redundant, unless RT data can be found that could not be explained by this new second-stage process alone. If the processing mechanisms of the feature-comparison model were to become more clearly specified, it might therefore become more similar to a network model.

It is important to evaluate current theories not only in terms of how well they handle available data, but with respect to their prospects for extension to the wider range of phenomena with which cognitive psychology must eventually deal. So far semantic-memory research has been almost exclusively concerned with meaning-comparison tasks involving just two words, sometimes with variations in the quantifier. The feature-comparison model can account for some of this data reasonably well. But if we take as our goal the description of the psychological representation of meaning, our theories must eventually analyze verbs, and allow us to represent the meaning of sentences of considerable complexity. Will it be possible to extend the feature-comparison model to accomplish this task?

There is reason to doubt it. To understand the apparent theoretical deficiencies of the feature-comparison model, it is helpful to distinguish between two aspects of meaning, which in linguistics are termed 'reference' and 'sense' (Frege, 1952). 'Reference' refers to the relationships between words and objects or events in the world, while 'sense' refers to the relationships of words to other words – that is, to their meanings within the linguistic system. Both of these aspects of meaning represent important problems for psychology, and in many respects they are clearly interrelated (e.g., in regard to the acquisition of word meanings). Nevertheless, the conceptual distinction between reference and sense is an important one for theories of semantic memory. The intuition underlying the feature-comparison model – that some instances are more 'typical' members of a category than others – essentially concerns referential meaning. However, the verification tasks that have provided the data base for semantic-memory models require the subject to rapidly compare word meanings – that is, to evaluate the 'sense relations' between words. The main structural assumption of the feature-comparison model, that word meanings are represented by sets of features weighted on their definingness, constitutes a hypothesis concerning the representation of sense relations.

The feature-comparison model has thus been directed at the problems of both reference and sense, although the distinction has not been kept clear. The marker-search model, on the other hand, is directly concerned only with sense relations. In the framework of the latter model, a major theoretical problem is to specify how a system of semantic markers can be mapped onto the perceptual system to account for people's basic ability to use words to refer to objects and events in the world. However, in the present paper we are solely concerned with evaluating semantic-memory models as accounts of the representation of sense relations. We will argue that the feature-comparison model lacks the theoretical power to represent a variety of concepts.

As a set-theoretic model, the feature-comparison model treats features as if they were independent 'units' of meaning. The decision process in both of the two stages in the feature-comparison model can be conceptualized as the 'summing up' of information from a set of independent comparisons between pairs of features. The principal theoretical device of this model – the weighting of features with respect to their definingness – thus makes it disturbingly similar to a class of weighted-feature models (perceptrons) that Minsky and Papert (1969) have proved to be in principle inadequate as theories of pattern recognition. The basic problem with such models is that they are unable to recognize visual properties such as connectedness, which depend on the relationships between features. If concepts in natural language can also be shown to be relational, this would suggest that perception-type models will also prove inadequate in the domain of language.

With respect to the representation of sense relations, a set-theoretic feature theory faces two hurdles – first, the problem of specifying a set of independent features; and second, the necessity to demonstrate that these features are sufficient to define concepts in natural language. It seems possible that these hurdles will prove insurmountable. A simple set-theoretic model can represent the relation of containment, as expressed in a sentence such as *A canary is a bird*, by assuming that the set of features defining the predicate is a subset of the features defining the subject. But how could it represent the relation of possession, which is expressed by a verb such as *has* (e.g., *The man has a car*)? This sentence does not refer to an entity that combines the features of *man* and *car*; nor does the sentence imply that *man* and *car* share some features. It does not seem possible to represent the abstract relationship expressed by *has* (or many other verbs, such as *can* and *does*) by combining sets of features using the simple boolean operations to which set-theoretic models have restricted themselves.

If the relation of possession cannot be represented by a set-theoretic feature model, then such models will also be unable to account for category membership, the problem which the feature-comparison model allegedly addressed directly. For example, instances of *money* are defined by their function as a medium of exchange. If we then analyze the concept 'exchange', we find it refers to an event in which possession of one object is superceded by possession of another. Then if we analyze the concept of 'possession', we find that it is similar or identical to the relationship expressed by the verb *has*, which it appears cannot be represented by a set-theoretic feature model. It follows that there can be no set of independent defining features for the category *money*; rather, membership in this, and many other categories (e.g., *toy, government, pet, game*) is defined by a relational decision rule. Such rules cannot be represented in terms of sets of independent features. Consequently, it appears that set-theoretic feature models, such as the feature-comparison model, are not sufficiently powerful to account even for category membership.

It might be argued that any theory of sense relations will ultimately be reducible to a theory of reference. Indeed, the initial plausibility of the feature-comparison model is largely dependent on the fact that it has been mainly applied to categories with clear perceptual referents, so that the postulated features have referential meaning. For example, one might suppose that many of the defining features of *canary* can literally be seen (e.g., 'yellow', 'small', 'flies', 'sings', etc.). However, Smith et al., have not answered the classic objections to the assumption that sense relations can be explained by a theory of reference (Frege, 1952; Katz, 1972). Accordingly, we presume that these objections stand against the feature-comparison model if it is considered to be such a theory.

It might also be asked, of course, whether network models can be extended to represent semantic relations other than containment (e.g., possession). While this is an open question, there is reason to believe they may. For instance, several network theorists (Anderson & Bower, 1973; Rumelhart, Lindsay & Norman, 1972; Schank, 1972) have used labeled relational arrows in order to represent different formal relationships between components. In linguistics, Katz (1972) has presented detailed proposals for representing the meanings of complex concepts in forms equivalent to graph structures. Thus while no network theory has yet been developed to deal with the full complexity of meaning, it at least may have the potential. Given the theoretical deficiencies of current set-theoretic models, it seems likely that if such models are elaborated they will become equivalent to a network representation.

But at the same time, it is clear that network models also face numerous serious difficulties in accounting for meaning, some of which have been mentioned in passing. The central issue is how network representations can be constrained. The distinction between set-theoretic and network models is, at heart, a difference in the conceptual resources that the two allow. Set-theoretic models are conceptually meagre and bare-boned, and as a result seem unable to capture the variety of tricks the human memory seems capable of performing. Network models face quite a different problem, for there are far fewer constraints on the kinds of representations they allow. The marker-search model has attempted to establish a few empirical constraints for a tiny fragment of English. While this attempt has met with some success, problems remain even here. If any such model is to be extended to other areas of language, it is essential that additional constraints be found.

This paper has focussed directly on the relative merits of two particular classes of psychological models of semantic memory. Nevertheless, we believe this discussion has significance for linguistics and philosophy as well, since it provides evidence that is inconsistent with some theories in these domains while demonstrating an empirical basis for others. If no psychological process model that can account for sentence verification results is consistent with the claim that truth is a continuous dimension, then a linguistic competence model based on that claim is unlikely to be justified. On the other hand, the psychological evidence that supports the marker-search model provides a broader empirical basis for the existence of those constructs (markers) that Katz (1972) needs to construct his definition of analyticity. The choice between models thus reflects not only narrow concerns within psychology, but also basic disagreements about the nature of human understanding.

REFERENCES

Anderson, J. R., & Bower, G. H. (1973) Human Associative Memory. Washington, V. H. Winston & Sons.

Bierwisch, M. (1969) On certain problems of semantic representation. *Found. Lang., 5,* 153–184.

Clark, H. H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. verb. Learn. verb. Beh., 12,* 335–359.

Clark, H. H., & Chase, W. G. (1972) On the process of comparing sentences against pictures. *Cog. Psychol.*, *3*, 472–517.
Collins, A. M., & Quillian, M. R. (1969) Retrieval time from semantic memory. *J. verb. Learn. verb. Beh.*, *8*, 240–248.
Frege, G. (1952) On sense and reference. In Geach, P., & Black, M. (Eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Oxford, Basil Blackwell & Mott.
Glass, A. L., & Holyoak, K. J. (1974) The effect of *some* and *all* on reaction time for semantic decisions. *Mem. Cog.*, *2*, 436–440.
Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974) Production frequency and the verification of quantified statements. *J. verb. Learn. verb. Beh.*, *13*, 237–254.
Holyoak, K. J., & Glass, A. L. (In press) The role of contradictions and counter-examples in the rejection of false sentences. *J. verb. Learn. verb. Beh.*
Katz, J. J. (1972) *Semantic Theory*. New York, Harper & Row.
Kintsch, W. (1972) Notes on the structure of semantic memory. In Tulving, E., & Donaldson, W. (Eds.), *Organization of Memory*. New York, Academic Press.
Lakoff, G. (1972) Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Papers from the Eighth Regional Meeting, Chicago Linguistics Society*. Chicago, University of Chicago Linguistics Department.
Landauer, T. K., & Meyer, D. E. (1972) Category size and semantic memory retrieval. *J. verb. Learn. verb. Beh.*, *11*, 539–549.
Loftus, E. F. (1973) Category dominance, instance dominance, and categorization time. *J. exp. Psychol.*, *97*, 70–94.
Meyer, D. E. (1970) On the representation and retrieval of stored semantic information. *Cog. Psychol.*, *1*, 242–300.
Meyer, D. E., & Schvaneveldt, R. W. (1971) Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *J. exp. Psychol.*, *90*, 227–234.
Minsky, M., & Papert, S. (1969) *Perceptrons*. Cambridge, The M.I.T. Press.
Posner, M. I. (1970) On the relationship between letter names and superordinate categories. *Q. J. exp. Psychol.*, *22*, 279–287.
Rabbitt, P. M. A. (1966) Errors and error correction in choice response tasks. *J. exp. Psychol.*, *71*, 264–272.
Rips, L. J., Shoben, E. J., & Smith, E. E. (1973) Semantic distance and the verification of semantic relations. *J. verb. Learn. verb. Beh.*, *12*, 1–20.
Rosch, E. R. (1973) On the internal structure of perceptual and semantic categories. In T. M. Moore (Ed.), *Cognitive Development and Acquisition of Language*. New York, Academic Press.
Rumelhart, D. E., Lindsay, P. H., & Norman, D. A. (1972) A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization and Memory*. New York, Academic Press.
Schaeffer, B., & Wallace, R. (1970) The comparison of word meanings. *J. exp. Psychol.*, *86*, 144–152.
Schank, R. C. (1972) Conceptual dependency: a theory of natural language understanding. *Cog. Psychol.*, *3*, 552–631.
Smith, E. E. (1967) Effects of familiarity on stimulus recognition and categorization. *J. exp. Psychol.*, *74*, 324–332.
Smith, E. E., Rips, L. J., & Shoben, E. J. (In press) Semantic memory and psychological semantics. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 8. New York, Academic Press.
Smith, E. E., Shoben, E. J., & Rips, L. J. (1974) Structure and process in semantic memory: a featural model for semantic decisions. *Psychol. Rev.*, *81*, 214–241.
Sternberg, S. (1969) The discovery of processing stages: extensions of Donders' method. In W. G. Koster (Ed.), *Attention and Performance II*. Amsterdam, North-Holland Publishing Company.
Wilkins, A. T. (1971) Conjoint frequency, category size, and categorization time. *J. verb. Learn. verb. Beh.*, *10*, 382–385.

Résumé

On discute le fait que les théories de la mémoire sémantique suivent une divergence parallèle à celle des controverses concernant la représentation de la signification. Le modèle de comparaison des traits (Smith, Shoben et Rips, 1974) applique la théorie linguistique de Lakoff (1972) pour prédire le temps de réaction dans la vérification des phrases, alors que le modèle de recherche des marques, décrit ici, utilise le type de représentation sémantique défini par Katz (1972) pour expliquer des donnés analogues. Les deux modèles sont décrits et leur portée est revue. Le modèle de recherche de marques se vérifie bien mais en revanche une prédiction majeure de modèle de comparaison des traits est infirmée.

On discute le fait que le modèle de comparaison des traits est inadéquat pour rendre compte de la représentation sémantique tant que sa conception des consituants sémantiques reste inchangée.