Psychological Review
1974, Vol. 81, No. 3, 214–241

STRUCTURE AND PROCESS IN SEMANTIC MEMORY: A FEATURAL MODEL FOR SEMANTIC DECISIONS [1]

EDWARD E. SMITH, EDWARD J. SHOBEN, AND LANCE J. RIPS [2]

Stanford University

A model is proposed to account for recent findings on the time needed to decide that a test instance is a member of a target semantic category. It is assumed that the meaning of a lexical term can be represented by semantic features. Some of these features are essential or defining aspects of a word's meaning (defining features), while others are more accidental or characteristic aspects (characteristic features). This defining versus characteristic distinction is combined with a two-stage processing mechanism in such a way that the first stage determines the similarity between the test instance and target category with respect to both defining and characteristic features, while the second stage considers only agreement between defining features. This model is shown to be consistent with most semantic memory effects, and two new experiments provide further detailed support for it.

As psychologists have become increasingly interested in the structure and processing of natural language, new areas of research have emerged. One such area—called semantic memory by Tulving (1972)—is primarily concerned with an individual's knowledge of the lexicon and how he utilizes this knowledge in understanding semantic relations. This concern with semantic memory has led to a series of experiments on semantic categorization. In these experiments, one decides whether a lexical item, for example, *robin,* is a member of a specified semantic category, for example, *bird* (Collins & Quillian, 1969). The times needed to make such decisions are of interest not only because they tell us something about semantic categorization processes but also because they provide information about the basic structure of semantic memory. In this article we will present a theory which accounts for much of the data on semantic categorization, but before doing so, we briefly consider some earlier models of semantic memory.

In discussing these earlier models, it is helpful to make a distinction between network and set-theoretic models (see Rips, Shoben, & Smith, 1973). The network models, exemplified by the work of Collins and Quillian (1969; 1972a) and Rumelhart, Lindsay, and Norman (1972), assume that words or their conceptual counterparts exist as independent units in semantic memory, connected in a network by labeled relations. As an example, the proposition *A robin is a bird* is represented in the Rumelhart et al. (1972) model as two nodes, corresponding to *robin* and *bird,* connected by the *is a* relation (that is, by the subset relation). Verification of this proposition requires retrieval of the stored relations between *robin* and *bird* and a comparison of these relations to that asserted in the proposition. While processing involves both retrieval and comparison operations, thus far these models have placed major emphasis on the retrieval component.

In contrast to theories emphasizing network structures and retrieval processes are those proposals which we term set-theoretic models. In these models concepts like *robin* and *bird* are represented by sets of elements. The elements might be unique descriptions

[1] The authors are indebted to R. C. Atkinson and E. A. C. Thomas for their comments on an earlier version of this article, and to Tom Schumacher for his careful assistance in executing the research reported here. We are particularly indebted to H. H. Clark for the advice he has given us in many stages of our work. The research was supported by U.S. Public Health Service Grant MH-19705.

[2] Requests for reprints should be sent to Edward E. Smith, Department of Psychology, Stanford University, Stanford, California 94305.

or images of exemplars, attributes, names of subsets or supersets, or some mixture of these types (see Meyer, 1970; Schaeffer & Wallace, 1970). For our purposes, the most important set-theoretic model is the one based on attributes. In this attribute model, *robin* and *bird* are represented by sets of their defining attributes, and verification of the proposition *A robin is a bird* is based on a comparison process that determines whether every attribute of *bird* is also an attribute of *robin*. Note that this model determines semantic classification correctly only if the attributes associated with *bird* are necessary and sufficient criteria for category membership. Other attributes which we characteristically associate with *bird* (e.g., that they can fly) cannot be a included in this set of attributes, for otherwise a proposition like *A penguin is a bird* will be incorrectly disconfirmed.

This attribute comparison process forms the second stage of a two-stage model proposed by Meyer (1970) to account for the verification of universally quantified propositions such as *All robins are birds*. In the first stage of Meyer's model, a set is retrieved which contains the names of all categories that have some members in common with the category of the predicate noun (*bird* in the previous example). Then this set of intersecting categories is searched for the subject noun. The order of search is determined by the set difference between each of the intersecting categories and that of the predicate noun, where the set difference between the predicate and an intersecting category is measured by the number of exemplars contained in the two categories that are not shared. If the subject noun is found in this set of intersecting categories, the second-stage attribute comparison is executed. Thus, structurally, the model represents each lexical item in two different ways—by the names of the item's intersecting categories and by the item's defining attributes—while processing involves comparison mechanisms that operate on this information. For statements without explicit quantifiers (e.g., *a robin is a bird*), Landauer and Meyer (1972) have proposed a model which is similar to the first stage of the model just described, although no explicit rule for search order is given.[3]

The theory which we develop in this article shares a set-theoretic structure with the last models described as well as with the proposals of Schaeffer and Wallace (1970). But the present model differs from this earlier work in several important ways. First, the present model represents semantic information by only one means—semantic features. Further, we consider a broader range of features than those which strictly define a category. Finally, with regard to processing assumptions, we propose a two-stage comparison process in which the relationship between the two stages is probabilistic, in contrast to the strictly additive stage model of Meyer (1970). These structural and processing assumptions are described in detail in the first two sections of this article. The model is then used to derive predictions about categorization times in a situation where a subject must rapidly decide whether a test item is a member of a particular target category. In the third section we deal with a problematic issue in recent research, the effect of varying the size of the target category on categorization time, and in the final section we present a quantitative test of the model.

STRUCTURAL ASSUMPTIONS AND RELEVANT EVIDENCE

Representation of Meaning

We begin with what is really a meta-assumption: The meaning of a word is not an unanalyzable unit but rather can be represented as a set of semantic features. Semantic features have been used to describe systematically many aspects of language, including semantic universals (Bierwisch, 1967), disambiguation (Katz & Fodor,

[3] We note that our contrast between network and set-theoretic models is intended primarily as an organizational device rather than as a sharp division between different classes of models. For example, a more recent version of Collins and Quillian's (1972b) theory can be considered a hybrid of network and set-theoretic assumptions, for semantic classification is sometimes accomplished by retrieving stored relations, while in other cases a property comparison must be made.

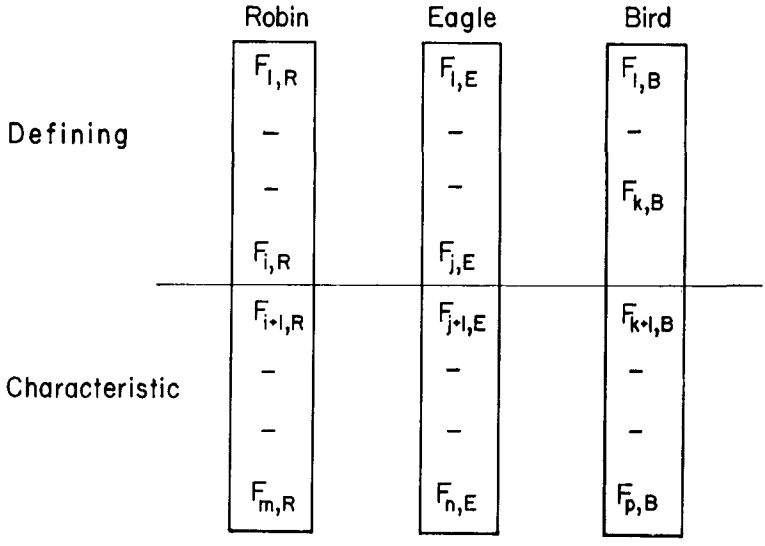

FIGURE 1. The meaning of a few lexical items in terms of discrete semantic features. (For any feature, the first subscript indicates its relative definingness, with lower numbers indicating greater definingness, and the second subscript specifies that it pertains to a particular item, with R standing for *robin,* E for *eagle,* and B for *bird.*)

1963), lexical acquisition (E. Clark, 1973), word association (H. Clark, 1970), reasoning (H. Clark, 1969), and word classification (Miller, 1969, 1972). Within this framework, we make the additional assumption that the features associated with a given category vary in the extent to which they define that category. This assumption is motivated by several facts. First, philosophers (e.g., Wittgenstein, 1953, Sections 65–80) have noted that few natural language categories have well-defined sets of necessary and sufficient criteria. Attempts by linguists to isolate such a restricted set in terms of a speaker's linguistic knowledge have proved problematic (see Bolinger, 1965). We believe that a solution to this problem is to think of a continuum along which some features will be more defining or essential aspects of a word's meaning, while others will be more accidental or characteristic features. To illustrate our assumption, which we will call the "characteristic feature" assumption, consider the word *robin.* Among the facts that an individual may know about this category are that robins are bipeds, have wings, have distinctive colors, and also that they perch in trees and are undomesticated. We propose that for most individuals, the first three of these features are considered more defining for the concept *robin* than are the last two, though an individual may be uncertain about which of these are strictly necessary *robin* features.

This variation in definingness must play a role in a categorization situation if accurate judgments are to be made. Specifically, the more defining features must be given greater emphasis. To make this notion of differential emphasis more explicit, we will later propose a processing model in which a subject in a categorization task initially considers all features, regardless of their definingness, and then restricts his attention to only those features that are above some lower bound for definingness. This bound therefore creates a distinction with features above the bound being considered defining features and those below the bound being termed characteristic. To illustrate with our example of *robin,* features that are concerned with being a biped, having wings, and having certain distinctive colors might all be considered defining, while features dealing with perching in trees and being undomesticated might be considered charac-

teristic. (See Bell & Quillian's, 1971, notion of criteriality for a related distinction.)

The characteristic feature assumption is meant to apply to superordinate terms, like *bird* and *animal,* as well. For example, part of the characteristic meaning of *bird* may be that it has a particular average size. To so concretize an abstract term like *bird* amounts to assuming that when we think of *bird* we have a typical bird in mind, an idea explicitly proposed by Rosch (1973, 1974). This aspect of the characteristic feature assumption, along with other aspects we have noted, are represented diagrammatically in Figure 1. Here the meaning of a few selected items in one semantic domain (birds) is depicted by lists of discrete features. The features for each item are presumed to be ordered, from top to bottom, by definingness, with an explicit bound creating the defining versus characteristic feature distinction. Figure 1 also depicts our second structural assumption: the number of defining features contained in an item's meaning decreases as the item becomes increasingly abstract; for example, *robin* contains more defining features than *bird.* This assumption has also been made by H. Clark (1970) and Meyer (1970), and it seems unavoidable when one views meaning in terms of defining features.

Since the characteristic feature assumption plays a major role in our theorizing, it is important to consider the evidence for it. There are two types of relevant evidence, linguistic analyses and experimental findings, and we consider them in turn.

Linguistic Evidence

The major linguistic evidence for the characteristic feature assumption comes from Lakoff's (1972) analysis of hedges, a class of modifiers whose major function seems to be that of qualifying predicates. Some sample hedges are listed in the first column of Table 1. Next to each hedge is given a subject noun-predicate noun pair with which the hedge can be appropriately used. That is, according to Lakoff, *A robin is a true bird, Technically speaking, a chicken is a bird,* and *Loosely speaking, a bat is a bird* are all acceptable sentences. But, as Lakoff notes, other combinations of the subject-predicate noun pairs with the hedges in Table 1 yield less acceptable sentences. For example, *Technically speaking, a robin is a bird* and *A chicken is a true bird* probably sound odd to most speakers of the language. Lakoff's intuitions about acceptability seem entirely reasonable to us, and they lead to the following question. Why is it that the hedge *a true* can be used with the *robin–bird* pair but not with the *chicken–bird* pair, while the reverse is the case for the hedge *technically speaking?*

TABLE 1

Hedges and Appropriate Subject–Predicate Pairs

Hedge	Subject noun–Predicate noun	Features of predicate noun shared by subject noun
A true	robin–bird sparrow–bird parakeet–bird	defining and characteristic
Technically speaking	chicken–bird duck–bird goose–bird	defining but not characteristic
Loosely speaking	bat–bird butterfly–bird moth–bird	characteristic but not defining

Lakoff's (1972) answer is in terms of defining and characteristic features. Since both *robin* and *chicken* meet the technical definition of a *bird,* there must be some other meaning difference between these two terms that determines their compatibility with the hedges, and this difference is one of characteristic meaning. The characteristic features of *bird* are also features of *robin* but not of *chicken.* Thus the hedge *a true* is used whenever the subject noun in a predicate nominalization contains both defining and characteristic features of the predicate noun, while *technically speaking* is employed when the subject noun contains the predicate noun's defining features but not its characteristic ones (see Table 1). In the same manner, the hedge *loosely speaking* signals possession of characteristic features but not defining ones. While there is far more to Lakoff's (1972) analysis of hedges than this, the above suffices to document a linguistic basis for the characteristic feature assumption. In addition, granting Lakoff's

TABLE 2

TYPICALITY RATINGS FOR INSTANCE–CATEGORY PAIRS

Birds		Mammals	
Instance	Rating	Instance	Rating
Robin	3.00	Deer	2.83
Sparrow	3.00	Horse	2.76
Bluejay	2.92	Goat	2.75
Parakeet	2.83	Cat	2.67
Pigeon	2.83	Dog	2.67
Eagle	2.75	Lion	2.67
Cardinal	2.67	Cow	2.58
Hawk	2.67	Bear	2.58
Parrot	2.58	Rabbit	2.58
Chicken	2.00	Sheep	2.58
Duck	2.00	Mouse	2.25
Goose	2.00	Pig	2.17

Note. Table 2 is adapted from "Semantic Distance and the Verification of Semantic Relations" by L. J. Rips, E. J. Shoben, and E. E. Smith, *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 1–20. Copyright 1973 by Academic Press. Adapted by permission.

arguments, one could then use hedges to determine characteristic features by noting, for example, what incidental features are common to those bird exemplars that take the hedge *a true*.

There is further linguistic evidence that superordinates contain characteristic meaning, specifically the analysis of relative adjectives given by Sapir (1944) and Bierwisch (1971). Basically they propose that when a relative adjective, like *big*, is applied to a particular superordinate term, it is understood in relation to the characteristic meaning of that superordinate. To illustrate, consider Sentence *a*:

(*a*) *That bird is big.*

In order to evaluate the truth of this sentence, one needs to compare the size of the designated bird with some size norm. Following Bierwisch's argument, this norm may often be the average size of the direct superordinate class, that is, the average size of a bird. Hence, Sentences *b* and *c* are paraphrases of *a*:

(*b*) *That bird is big for a bird.*

(*c*) *That bird's size is bigger than the average size of a bird*

Thus a sentence like *a* contains an implicit comparison to a particular dimension value that characterizes the class to which the subject noun belongs, and this comparison is made explicit in *c*. Phrased somewhat differently, the superordinate class contains characteristic features (e.g., a particular average size), and constructions like *a* are understood in relation to these characteristic values.

Experimental Evidence

The major experimental evidence for our characteristic feature assumption comes from studies where subjects are presented a set of instance–category pairs (e.g., *robin–bird*), with several instances being used for each category, and are simply asked to rate how typical each instance is of its associated category (e.g., Rips et al., 1973; Rosch, 1973). Sample typicality ratings from Rips et al. (1973) are presented in Table 2 for instances of the bird and mammal categories. (For these ratings, the subjects were actually asked to rate the relatedness for each instance–category pair rather than its typicality, but Rips et al. showed that relatedness and typicality ratings for these items correlated about .90.) The first thing to note in Table 2 is that there are differences in how typical various instances are judged to be of their superordinate, and these differences are highly replicable (Rosch, 1974). Moreover, the typicality ratings are in excellent congruence with the results of Lakoff's (1972) hedge analysis, in that instances rated as typical (e.g., *robin* and *sparrow* for the bird category) are those that take the hedge *a true*, while instances judged to be atypical (e.g., *chicken* and *duck*) go with *technically speaking*. In light of this, a simple explanation of the typicality ratings is that such ratings primarily reflect the extent to which the characteristic features of a superordinate are similar to the features of an instance. The defining features of a superordinate cannot exercise much influence on typicality ratings because all instances contain these features. Thus typicality ratings are readily interpretable in terms of our characteristic feature assumption and, in fact, provide an excellent

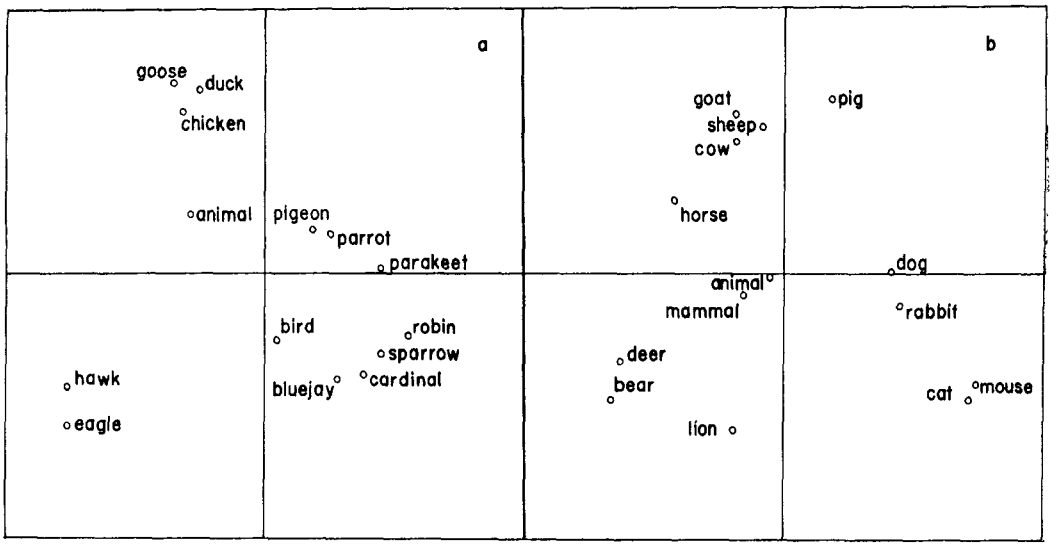

FIGURE 2. Multidimensional scaling solutions for birds (a) and mammals (b). From "Semantic Distance and the Verification of Semantic Relations" by L. J. Rips, E. J. Shoben, and E. E. Smith, *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 1–20. Copyright 1973 by Academic Press. Reprinted by permission.

means of determining the extent to which an instance and a category share characteristic features.

So we are led to the proposal that the rated typicality of an instance–category pair reflects the extent to which the two lexical items differ on some basic characteristic features. Given this, the ratings for any category might be represented in a multidimensional space, where the dimensions of the space would be interpreted as the characteristic features of the category and the distance between an instance and its category depicts the dissimilarity between them with respect to these features or dimensions. Rips et al. (1973) obtained such multidimensional representations of typicality ratings for birds and mammals, and their results are presented in Figure 2. The multidimensional scaling procedure used was Carroll and Chang's (1970) INDSCAL program which takes as input the similarity ratings of all possible pairings of the relevant items and yields as output a Euclidean solution in n dimensions. Two-dimensional solutions suffice for birds and mammals, and the distance measure seems to capture a good deal of the typicality ratings in Table 2, for example, the distance between *robin* and *bird* is less than that between *chicken* and *bird*.[4]

One important aspect of this scaling analysis is that the dimensions of the spaces presumably reflect underlying characteristic features of the category and insofar as these dimensions are interpretable, one may gain some insight into the nature of these features. Rips et al. (1973) have offered an interpretation of the dimensions of the spaces in Figure 2. For the birds solution (Figure 2a), the dimension spanning the horizontal

[4] We are assuming here that differences among characteristic features are combined via a Euclidean metric rather than by some other metric. This assumption is supported by some of our unpublished results on the scaling of bird and mammal instances. Using the nonmetric MDSCAL program, developed by Shepard (1962a, 1962b) and Kruskal (1964), we found that non-Euclidean solutions were less appropriate than the Euclidean solution. Specifically, the minimum stress value, a measure of goodness of fit, was less for the Euclidean solution (.190 for birds and .188 for mammals) than for either the City Block metric (.218 for birds and .197 for mammals), or the Dominance metric (.232 for birds and .236 for mammals). The latter two metrics are the ones most often considered as viable alternatives to the Euclidean combination rule (see Cross, 1965).

axis was interpreted as a size continuum with *hawk* and *eagle* at one end and *robin* at the other, while the vertical dimension seemed to order the birds in terms of their predatory relations with game birds at one end and predators at the other. This second dimension has been referred to as predacity (Rips et al., 1973) or ferocity (Henley, 1969). For the mammals space (Figure 2b), the same two dimensions emerge with the horizontal dimension again reflecting size (*deer* and *bear* at one extreme and *mouse* at the other) and the vertical dimension offering something of a predacity continuum (with wild animals and farm mammals being found at the two extremes). Our interpretation of the dimensions is strengthened by the fact that a metric program (INDSCAL), which yields a unique orientation of the axes, was used, and by the fact that Henley's (1969) scaling of mammals produced the same two dimensions.

Thus far the experimental evidence for our characteristic feature assumption has centered on rating data. There are, however, other findings which conform to our assumption. Consider a recent study by Rosch (1974) on typicality and sentence acceptability. First, Rosch gave one group of subjects superordinate terms, like *bird,* and had her subjects generate a sentence for each superordinate. A sample construction might be, *The tree has about twenty birds perched in it.* Rosch then substituted an instance for the superordinate term in the constructed sentence, where this instance varied in how typical it was of the superordinate (e.g., *robin* or *chicken* might be substituted for *bird* in the above example). Next, Rosch gave the altered sentences to a second group of subjects who rated each sentence for acceptability. The major finding was that rated acceptability decreased as typicality decreased. The reason for this seems to be that the sentence originally generated by the first group had a congruence between the characteristic features of the superordinate (e.g., *bird*) and the predication in the sentence (e.g., *perch in trees*), and this correspondence was violated when an atypical instance was substituted because an atypical instance shares few characteristic features with its superordinate. Indeed, Rosch (1974) notes that in some cases the feature correspondence violated by atypical instances involved the features of size and predacity, which are, of course, the characteristic features identified by the scaling results of Rips et al. (1973).

Implications for Alternative Models

To summarize, there is a sound linguistic basis for our characteristic feature assumption, and experimental findings dealing with typicality (ratings, multidimensional scaling, and sentence substitutions) are readily interpretable in terms of this assumption. Moreover, the evidence presented thus far is somewhat problematic for previous models of semantic memory. First, consider network models in relation to the results on hedges and typicality. These results point to the conclusion that category membership is a matter of degree, with typical instances being better members than atypical ones, and this conclusion is not in keeping with the salient, structural aspects of network models. Thus in a network structure like that of Collins and Quillian (1969) or Rumelhart et al. (1972), each instance is presumably connected to its direct superordinate by a single labeled relation. Such a structure can only account for typicality effects in a manner that we consider to be ad hoc. For example, it might be assumed that subset relations for atypical instances (e.g., the relation between *chicken* and *bird*) involve intermediate nodes (e.g., *chicken* is a subset of *fowl*, which in turn is a subset of *bird*), while subset relations for typical instances are more direct. Alternatively, a network model could account for typicality effects by simply positing that the subset relations for typical instances are stronger or more accessible than those for atypical instances. However, both of these modified network models face considerable difficulties, and some of these problems have been detailed elsewhere (Rips, et al., 1973; Smith, Rips, & Shoben, 1974).

Two problems with these modified network models have not been mentioned previously, and they deserve some comment here. First, there is the question of how

these models can handle the following situation: One instance (e.g., *robin*) is judged to be more typical than another (e.g., *chicken*) of a given superordinate (e.g., *bird*), but these typicality judgments reverse for a higher order superordinate (e.g., *chicken* is more typical of *animal* than is *robin*). In an experiment to be described below, we report (Experiment 2) just such an interaction between items and category size. To see the problem for modified network models, note that if intermediate nodes or inaccessible relations exist between *chicken* and *bird,* then, by standard assumptions, these same nodes or relations must be interposed between *chicken* and *animal.* This, in turn, means that *chicken* must be less typical of *animal* than is *robin.* The second problem to be considered here is that the modified network models have difficulty with certain types of hedges like *loosely speaking* (*Loosely speaking, a bat is a bird*) *or fake* (*A decoy is a fake bird*). For although both propositions would be judged as true, they could not be verified by retrieving a subset relation between the two pairs of nouns since no such relations exist. To circumvent this problem, one might posit that a superordinate like *bird* is represented by several different nodes in the network, one node representing the strict or literal meaning, a second representing a loose meaning of *bird,* and a third representing a class of pseudoinstances. But this solution seems unparsimonious at best.

A similar state of affairs obtains when we consider the typicality and hedge results in relation to the two-stage set-theoretic model of Meyer (1970) or the related model of Landauer and Meyer (1972). In these models one would have to attribute typicality effects to variations in the order in which intersecting categories are searched. Since Meyer's model already contains a determinant of search order—set differences (see Introduction)—the relations between this factor and typicality would have to be explicated. Further, the same difficulties arise here as for the network models. Specifically, in Meyer's model the size of set differences could not explain an interaction between items and category size, like that mentioned previously. The reasoning here is that, if the set difference of *chicken* and *bird* is large relative to that of *robin* and *bird,* then it follows that the set difference of *chicken* and *animal* must be larger than that between *robin* and *animal.* Also, problems again arise with hedges like that in *Loosely speaking, a bat is a bird* since there is no intersection between the categories *bat* and *bird,* and hence there is no means for confirming this proposition. Once more we could invoke the assumption that concepts are represented as several distinct categories, but this may lead to a proliferation of categories.

In general, then, alternative models, as currently stated, encounter difficulties in explaining the phenomena of typicality and hedges. Their difficulties are due to their failure to include structural assumptions that would make category membership a matter of degree. While we have given some examples of how these models might be modified structurally so as to reduce these difficulties, the feasibility of the modifications is placed in doubt by the foregoing considerations. In contrast to this state of affairs is the present structural approach, depicted in Figure 1. This representation of concepts contains characteristic as well as defining features, and consequently, it is clearly compatible with the notion that category membership is a matter of degree. In this respect, we consider it a marked advance over alternative representations.

A PROCESS MODEL FOR SEMANTIC CATEGORIZATION

Task Description

We now have a structure for semantic memory, and it remains to specify how it may be processed to yield answers in a task requiring semantic classifications. These tasks all require a subject to determine on each trial whether a test instance (e.g., *robin*) is a member of a target category (*bird*) or not. However, the format of this basic task may vary considerably. Thus a subject may be presented with sentences of the form *An S is a P* or *All S are P,* where S and P denote test instance and target

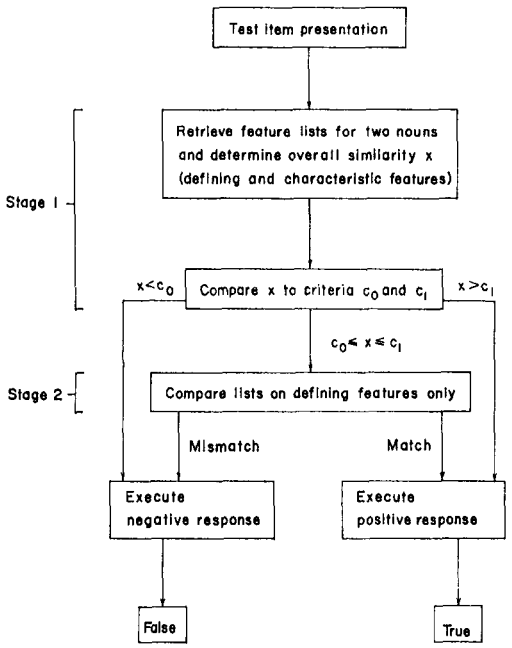

FIGURE 3. A two-stage, feature comparison model for semantic categorization tasks. (See text for explanation.)

category, respectively, and the subject is then required to determine the truth or falsity of the test sentence as quickly as possible. In another variant of the basic task, a target category is specified at the start of each trial, and a subject decides whether a test instance is a member of that category (*true* response) or not (*false* response). In what follows, we ignore distinctions between these different task variants. In all of these tasks, errors are usually infrequent and major interest centers on the subject's reaction time to respond *true* or *false* correctly.

Processing Assumptions

Before specifying our processing assumptions in detail, we must define more precisely what we mean by a semantic feature of a lexical item. We assume that a lexical item is represented by a set of relevant semantic dimensions (e.g., size).[5] Each dimension is associated with a weight which indicates how essential that dimension is in the definition of the lexical item, dimensions with high weights being more defining than those with low weights. For each item, there is a distribution of possible values on each relevant dimension, where this distribution corresponds to one's subjective impressions of the relative frequencies of the dimension values for that item. At any given time, the meaning of a lexical item can be represented as a list of values (with associated weights), one value being sampled from each dimension. These momentary values are what we mean by features.

Our processing assumptions (based, in part, on a model by Atkinson & Juola, 1973, 1974) are embedded in the multistage model presented in Figure 3, which we shall refer to as the "feature comparison" model. This model represents a more refined statement of the feature comparison process described in Rips et al. (1973). The basic idea is that a semantic categorization may require two distinct comparison stages. The first comparison stage includes three processes. First, lists of features for the instance and category are retrieved, including features drawn from characteristic as well as defining dimensions. Next, these two lists are compared, with respect to all features, yielding a measure, x, of overall similarity. This measure will be treated as a random variable for a given instance–category pair. The measure takes into account the proportion of the category's dimensions which are shared by the test item and the proximity of values (features) on each of these common dimensions. (At this point, it is difficult to be more specific about the exact combination rule used to compute overall similarity, but see Footnote 4.) Finally, the resulting x is compared to two criterial levels of overall similarity, preset before the start of the trial, one, a high level called c_1, and the other, a low level called c_0. If x exceeds c_1, then a positive response (*true*) can be made, while if x is less than c_0, a negative response (*false*) is executed.

[5] The use of the term "dimension" suggests a continuum of values. It seems to us that positing too many values may be less confining than positing too few, as when linguists treat all defining features as binary valued. On this point see Rosch (1973, 1974).

When x falls in the intermediate range between c_0 and c_1, a second comparison stage is necessary before a response can be determined. For both the category and the instance, this second stage separates the more defining features from the characteristic ones on the basis of feature weights and then compares the set of defining features of the category to those of the test instance. A positive response can be made if (*a*) each defining dimension of the category is also a defining dimension of the instance, and (*b*) the particular values (features) on these dimensions which the instance possesses are within the range of allowable values for the category. Note that Condition *b* is necessary to insure that one does not respond *true* to *A mother is a father,* where the instance and category share all defining dimensions.

To illustrate how this feature comparison model works, consider what happens when the statement *A robin is a bird* is presented for verification. First, the lists of defining and characteristic features for *robin* and *bird* are retrieved, and then the lists are compared to determine overall semantic similarity. We have noted that there is some trial-to-trial variation in the actual features that enter into this comparison, and consequently the overall similarity of *robin* and *bird* can be described by some distribution of x values. Since *robin* and *bird* are presumably very similar with respect to characteristic (as well as defining) features, most of the relevant distribution will contain relatively high values. Hence there is substantial probability that the value, x, obtained on any trial will exceed c_1 and that no second-stage processing will be needed. This omission of the time-consuming second stage will result in a fast, positive reaction time. On those trials where x is less than c_1 but greater than c_0, second-stage processing will occur. Assuming this second stage is relatively error free, a positive response will be made, and, of course, it will have a longer reaction time than the previously considered case where there was no need for second-stage processing. To round out this description, we should also note that on some small percentage of the trials, x may be less than c_0, and the subject will erroneously make a negative response to the *robin–bird* pair.

What the above indicates is that the average correct *true* reaction time (over trials) in a semantic categorization task is a mixture of two types of responses, fast times based on single-stage processing and slower times that result from dual-stage processing. Variations in *true* reaction times are assumed to be due to differences in this mixture. That is, if, on the average, one item results in a faster *true* reaction time than a second item, then the first item had more single stage, fast times and fewer dual stage, slow times than the second item. Exactly the same logic is used in dealing with correct *false* reaction time. Fast *false* times occur when the overall similarity between the test instance and category is so low as to be less than c_0, while slower *false* times occur when x is in the intermediate range and second-stage processing must be used. Thus variations in correct *false* times are also due to variations in the proportions of the two types of responses in a mixture.

Before deriving explicit predictions, we should comment on some general aspects of the feature comparison model. Note that the first stage of the model may be characterized as holistic, in that it considers all features, and as intuitive, in the sense that it considers only overall similarity of features rather than which features are similar. Also, the first stage is error prone, in that for *true* items x may occasionally be less than c_0, resulting in an erroneous *false* response. In contrast, the second stage is selective as it considers defining features, logical in that it bases its decision on a procedure that evaluates only defining features, and relatively error free. This contrast between early holistic and later analytic processing accords well with our introspections that decisions about logical matters are sometimes made quickly on a basis of similarity, while at other times decisions are the result of a more deliberative process.[6]

[6] Our assumption that the first comparison stage is holistic, in that the overall similarity of features is considered, receives support from the find-

Basic Predictions and Results

The predictions that will be considered in this section are those that follow most directly from our assumptions. It is worth noting that the predicted effects, particularly those pertaining to *false* reaction time, are among the largest effects reported in the semantic memory literature.

The first strong prediction is that for any target category, *true* reaction time should decrease as the typicality of the test instance increases. The rationale is as follows. All instances of the target category share roughly the same number of defining features with that category, but the number of shared characteristic features between instance and category increases with the typicality of the instance. Hence, as typicality increases, x increases. This in turn increases the probability that $x > c_1$, increasing the proportion of reaction times based only on single-stage processing. This typicality prediction has now been confirmed in several studies. Rips et al. (1973) and Rosch (1973) both obtained typicality ratings for instance–category pairs for different categories and then showed that *true* reaction time decreased with rated typicality. In the latter study, the difference between typical and atypical instances was on the order of 60 milliseconds. Similarly, Smith (1967), Wilkins (1971), and Loftus (1973) showed that the *true* reaction time to an instance–category pair was faster if that instance was more likely to be given as a response to that category in norms like those of Cohen, Bousfield, and Whitmarsh (1957) or Battig and Montague (1969). Since such category norms correlate very highly with typicality ratings (Rips et al., 1973) the results of Smith, Wilkins, and Loftus may be taken as three more demonstrations of the typicality effect. In these three studies the size of the effect ranges from about 50 to 150 milliseconds. It is worth emphasizing that while Wilkins assumed that category norms measure the co-occurrence or conjoint-frequency of an instance-category pair in the language, we are explicitly assuming that such norms reflect typicality. Category norms may reflect typicality either because the subjects contributing to the norms based their responses mainly on typicality, or because the subjects based their responses on conjoint-frequency, and typicality is a major cause of variations in conjoint-frequency (see Smith et al., 1974).

For correct *false* reaction times, the basic prediction from the feature comparison model is again concerned with the semantic similarity of the instance–category pair presented. When dealing with *false* items, we shall refer to similarity variations as variations in semantic relatedness (rather than typicality), and the basic prediction is that correct *false* reaction time should decrease as relatedness decreases. The rationale is similar to that used with *true* items. A decrease in the semantic relatedness of an instance–category pair in a *false* item indicates a decrease in the number of shared defining and characteristic features. Hence, as relatedness decreases, x decreases, and this increases the probability that $x < c_0$. This in turn increases the proportion of fast, correct *false* reaction times based on single-stage processing.

This relatedness effect has been documented in four recent experiments (Collins & Quillian, 1970; Meyer, 1970; Rips et al., 1973; Wilkins, 1971). In the Collins and Quillian experiment, a related *false* pair was one in which the test instance was an instance of a relatively direct superordinate of the target category. For example, if *animal* was the target category, then *tree* would comprise a related *false* instance, (since *tree* and *animal* share a direct superordinate, *living thing*), while *magnesium* would constitute an unrelated *false* instance. In terms of the feature comparison model, the test instance and target category were semantically more similar in the related than the unrelated condition, since there are clearly more shared defining features in related than unrelated *false* pairs. Conse-

ing that the Euclidean metric provides a better scaling solution for birds and mammals than non-Euclidean metrics (see Footnote 4). The importance of this finding for our present concern is that there is evidence that the Euclidean metric is associated with holistic or integral dimensional structures (Garner & Felfoldy, 1970; Shepard, 1964).

quently, correct *false* reaction time should have been faster for the unrelated than the related pairs. The Collins and Quillian results confirm this prediction, with the relatedness effect ranging from 100 to 140 milliseconds. Wilkins' (1971) experiment was basically the same as that of Collins and Quillian, and he replicated their relatedness effect, though the magnitude of Wilkins' effect (40 milliseconds) was considerably smaller than that obtained by Collins and Quillian. Rips et al. (1973) varied relatedness in the same way as Collins and Quillian and Wilkins, and obtained the same relatedness effect; in the Rips et al. experiment the effect was on the order of 250 milliseconds. Finally, Meyer's (1970) manipulation of the set-theoretic relation in *false* items can be interpreted as a relatedness variation. Specifically, Meyer used three types of *falses* that may be classified according to the set-theoretic relation of the instance to the category: (*a*) superset relation, for example, *All stones are rubies,* (*b*) overlap relation, for example, *All mothers are writers,* and (*c*) disjoint relation, for example, *All houses are vacuums.* Clearly, as one moves from superset to overlap to disjoint relations, there is a decrease in relatedness which is due to the decrease in defining and characteristic features shared by test instance and target category. And, as expected by the current model, correct *false* reaction time decreased from superset to overlap to disjoint statements, the range of this effect being 185 milliseconds.

The above indicates that most existent data on *false* reaction time can be explained in terms of variations in the probability that only first-stage processing is needed. Or to put it another way, most *false* reaction time findings can be explained without recourse to a detailed examination of second-stage processing. However, some very recent findings by Glass, Holyoak, and O'Dell (1974) indicate that this explanation may be incomplete. These authors present data that suggest there may be some effects on *false* reaction time that can only be accounted for by variations in second-stage mechanisms. This in turn suggests that further developments of the feature comparison model will need to include a more detailed analysis of second-stage processing.

Implications for Alternative Models

All in all, the basic predictions from the feature comparison model—*true* reaction time decreases and *false* reaction time increases with shared features (overall semantic similarity)—have solid empirical support. Again, it is of interest to see how well other models of semantic memory can handle these results. The difficulties that these models have in accommodating typicality effects have already been discussed in relation to typicality ratings (see p. 220), and the same problems arise when typicality effects are manifested in reaction times. With regard to the relatedness effect on *false* latencies, Collins and Quillian (1972a) have explored one way in which a network-retrieval model might explain such an effect. Their notion is that when a *false* pair is presented, other extraneous network pathways may be activated, and more extraneous paths will be activated for a related than an unrelated pair. If these paths have to be checked, then *false* reaction time should increase with relatedness. The assumption that a related pair contains more semantic connections than an unrelated pair seems close to our own proposal that related pairs share more semantic features than unrelated pairs. However, Collins and Quillian have not been completely explicit about the type of extraneous paths that must be considered. Meyer's (1970) treatment of the relatedness effect also bears a resemblance to ours. In Meyer's model, disjoint statements, which are often unrelated, are disconfirmed solely on the basis of first-stage processing, while the more related overlap and superset statements require second-stage processing and, hence, longer latencies. However, Meyer's model cannot make predictions for relatedness variations in disjoint statements. Such an effect, amounting to a difference of 360 milliseconds, has been obtained by Rips et al. (1973). We note, however, that Meyer (1970) does provide some discussion of this problem. He suggests that the first stage of his model might take into account the association strength

between the subject and predicate nouns, rather than the set difference between these nouns. Assuming that association strength is similar to semantic relatedness, Meyer's suggestion seems to lead to a model close to the one proposed here.

Thus, the current feature comparison model seems to have some advantage over earlier models in explaining the effects of semantic similarity on *true* and *false* reaction time. Moreover, the current viewpoint has also introduced some parsimony into the empirical literature, in the sense that the conjoint-frequency effect on *true* reaction time (Wilkins, 1971) may be reducible to typicality and that the set relation effect on *false* reaction time (Meyer, 1970) seems to be, at least in part, another manifestation of relatedness. Further, as has been noted, many of the findings covered in this section are among the most sizeable in the literature. But despite the importance of these findings they have not been the central focus of interest in semantic memory research. Rather, this favored role has fallen to the category size effect, to which we now turn.

ISSUE OF CATEGORY SIZE

Conceptions of the Category Size Effect

In semantic memory research the most common variable has been the size of the target category, where category size refers to the number of category members. The most popular way of manipulating this size variable has involved the use of nested triples of categories, for example, *robin, bird,* and *animal.* To illustrate the procedure, the subject may be asked on one trial to determine whether *A robin is a bird* and on another whether *A robin is an animal.* Since *bird* is nested within *animal,* the latter category is the larger in a logical sense and presumably in a psychological sense as well. When category size is manipulated in this fashion, numerous studies have found that *true* reaction time increases with category size (e.g., Collins & Quillian, 1969; Meyer, 1970); others have shown that *false* times also increase with category size (e.g., Landauer & Freedman, 1968; Meyer, 1970). In what follows, we will place major emphasis on the *true* times effect.

The reason for this interest in the category size effect on *true* reaction time is certainly not the magnitude of this effect, as it is often less than 50 milliseconds (see, e.g., Landauer & Meyer, 1972; Wilkins, 1971). Rather the effect is of interest because it has played a central role in the development of current semantic memory models. Consider first the network model of Collins and Quillian (1969, 1972a). While category size per se is of no consequence to this model (Collins & Quillian, 1970), the variation due to the use of three nested categories is a critical one. That is, Collins and Quillian (1969, 1972a) explicitly assume that while an instance (*robin*) is directly connected to its immediate superordinate (*bird*), this same instance may only be indirectly connected to a remote superordinate (*animal*) via an intermediate node (in this case *bird*). Consequently, verification of *A robin is a bird* requires retrieval of only one relation, whereas verification of *A robin is an animal* may require retrieval of two relations, and thus there should be an apparent category size effect. We note that while this effect is not an obligatory prediction of the model (see Collins & Quillian, 1969, p. 242), it is the major finding that has been offered in support of the model. The category size effect also follows quite naturally from Meyer's (1970) two-stage model. Since sets close in size to the category will be the earliest ones considered in the first stage, the entire process will usually terminate faster for the *robin–bird* pair than the *robin–animal* pair, as *robin* is closer in size to *bird* than to *animal.* Similarly, the category size effect has been offered as evidence for the model of Landauer & Meyer (1972). Here the effect arises because the larger the target category, the longer it will take to find the test instance among its contents. Again, we note that the category size effect is not a necessary prediction from the model in question, since a test instance may be located early in the search of a large category though very late in the search of a small category. Nevertheless, this category size effect is frequently cited as support for this type of model. Thus to undermine the category size effect is to undermine

some important evidence for alternative models, and we now proceed to challenge the generality of this critical effect.

In terms of the feature comparison model, an increase in category size for a *true* statement should affect both stages. Consider first the effect on the second stage that involves defining features. As noted earlier (Representation of Meaning), we assume that the number of defining features decreases as the term becomes increasingly abstract. Since, in the case of nouns, abstractness is positively correlated with category size, an increase in the size of the target category leads to a decrease in the number of defining features. This in turn implies a decrease in the duration of the second stage. This is so because all serial comparison models and most parallel ones predict that total comparison time is a function of the number of relevant components (e.g., Egeth, 1966; Townsend, 1971). Hence, if we consider only the second stage, an increase in category size should lead to a decrease in *true* reaction time.

But now consider the category size effect on the first stage. As one increases category size, there might be a substantial change in the number of characteristic features shared by the instance and category, and therefore a substantial change in the overall similarity which determines whether the second stage will be executed. The obvious question is, what is the direction of this change—does the increase in category size result in an increase or a decrease in overall similarity? If, on the one hand, it is an increase, then there will be a decrease in the probability that the second stage is executed, and this should speed *true* reaction time. In this case, category size affects both stages in the same way—both stages are facilitated—and the clear prediction would be that increasing category size shortens *true* times. On the other hand, if an increase in category size decreases overall similarity, then there will be an increase in the probability that the second stage is needed, and this should slow *true* time. In this case, category size affects the two stages differentially—the second stage is shortened, but it is also more likely to occur—and a clear prediction cannot be made.

In light of the above, we suggest that the reported increases in *true* reaction time with category size mean that increasing category size usually leads to a decrease in overall similarity and that the slowing of the latencies due to this decrease outweighs the facilitative effect of category size on the second stage. This explanation is obviously ad hoc. What is needed to make it persuasive is some demonstration that our model can lead to a correct prediction about category size that cannot be readily encompassed by models like those of Collins and Quillian (1969) and Meyer (1970). Such a prediction is: In those cases where increasing category size actually increases overall similarity, there should be a decrease in *true* reaction time, since the category size effects now facilitate both stages. This prediction has been verified by Rips et al. (1973). Rips et al. studied verification of instance–category statements where (*a*) bird instances were used with the target categories *bird* and *animal*, and (*b*) mammal instances were used with the target categories *mammal* and *animal*. Independent ratings of instance–category similarity showed that for bird instances the increase in category size from *bird* to *animal* decreased overall similarity, while for mammal instances the increase in size from *mammal* to *animal* increased overall similarity. Consistent with the aforementioned prediction, as category size increased, *true* reaction time increased for birds but actually decreased for mammals.

This finding of an inverse category size effect is certainly at odds with the emphasis of the theories of Collins and Quillian (1969), Meyer (1970), and Landauer and Meyer (1972). But there is a severe limitation on this finding. It has been obtained with only one category, *mammals*, and the result may reflect some peculiarity of this category. Hence there is a clear need to demonstrate that the finding in question holds for other categories. The following experiment provides this demonstration.

TABLE 3

MEAN CORRECT REACTION TIME (IN MILLISECONDS) AS A FUNCTION OF CATEGORY SIZE, SEPARATELY FOR TRIPLES IN WHICH THE SMALLER CATEGORY WAS THE MORE SIMILAR AND FOR TRIPLES IN WHICH THE LARGER CATEGORY WAS THE MORE SIMILAR

Smaller category more similar—Set 1			Larger category more similar—Set 2		
Instance	Smaller category Larger category	Reaction time	Instance	Smaller category Larger category	Reaction time
Butterfly	Insect Animal	1,077 1,325	Aluminum	Alloy Metal	1,267 1,144
Collie	Dog Animal	969 1,117	Cantaloupe	Melon Fruit	1,174 974
Copper	Metal Mineral	977 1,253	Cathedral	Church Building	1,027 1,167
Copperhead	Snake Reptile	1,083 1,398	Chimpanzee	Primate Animal	1,298 1,017
Daisy	Flower Plant	1,036 996	Coca-Cola	Pop Drink	1,009 897
Door	Entrance Opening	1,103 1,081	Diamond	Jewel Stone	1,135 1,101
Lemonade	Drink Liquid	1,022 1,041	Drum	Percussion instrument Musical instrument	1,292 1,095
Minute	Unit of time Unit of measurement	1,277 1,546	Fork	Silverware Utensil	1,122 1,023
Pear	Fruit Food	889 1,164	Guitar	Stringed instrument Musical instrument	1,089 1,042
Potato	Vegetable Food	1,058 983	Harvard	University School	978 1,000
Sparrow	Bird Animal	975 1,312	Scotch	Liquor Drink	1,043 988
Toe	Part of foot Part of body	1,235 1,172	Topaz	Gem Stone	1,143 1,130
Willow	Tree Plant	1,065 1,290	Wool	Cloth Material	921 1,170
Mean	Smaller category Larger category	1,059 1,206	Mean	Smaller category Larger category	1,115 1,058

A Study of Category Size: Experiment 1

We first need a set of ratings that establish whether an increase in category size leads to an increase or decrease in overall similarity for instance–category pairs. Loftus and Scheff (1971) have recently published a set of norms that can be used for this purpose. Loftus and Scheff obtained the norms by instructing a group of subjects to list superordinates for each of 50 instances. Production frequencies were then tabulated. We assumed that, for each instance, the production frequency of a category was a measure of that category's overall semantic similarity to the instance. Then we proceeded as follows. We considered only

those instances for which the two most similar categories were nested (e.g., the categories *flower* and *plant* for the instance *daisy*) and selected 13 triples (instance plus two categories) in which the more similar category was the smaller one, as well as 13 triples in which the more similar category was the larger one. These 26 triples, which form a representative sample of the Loftus and Scheff norms, are presented in Table 3. For each triple, two statements of the form *An S is a P* were typed on separate cards, one containing the instance and the smaller category and the other, the instance and the larger category. Then all 52 of these *true* statements were randomized with an equal number of *false* statements, constructed by pairing instances and categories from different triples, and were presented one at a time to a subject for rapid verification. Thirty young adults from the Stanford community served as subjects.

The crucial prediction is that, for the triples in which the smaller category was the more similar one (Set 1 triples), *true* reaction time should increase with category size, while for those triples where the larger category was more similar (Set 2), *true* reaction time should decrease with category size. The data, presented in Table 3, indicate that this prediction was confirmed. For the Set 1 triples, nine triples show the usual category size effect, while for Set 2 triples, ten show the reverse category size effect. To assess the results statistically, an analysis of variance was carried out in which category size and type of triple (Set 1 versus Set 2), were fixed effects with subjects and triples-within-sets as random effects (H. Clark, 1973). The use of two random effects in this analysis necessitated estimation of the F ratios (symbolized F') for main effects of the fixed factors and their interaction. This was accomplished by the quasi-F procedure recommended by Winer (1971, p. 377). There was no significant overall effect of category size ($F' = 2.51$, $df = 1/29$, $p > .10$), nor was there any difference between the mean reaction times for Set 1 and Set 2 triples ($F' = 1.37$, $df = 1/29$, $p > .25$). However, the interaction of these two factors was highly significant ($F' = 10.32$, $df = 1/38$, $p < .005$), where this interaction, of course, reflects the critical prediction. This interaction was also reflected in the error data. For Set 1 triples, smaller categories resulted in a lower error rate (7.2%) than larger categories (8.7%), while for Set 2 triples, larger categories produced a lower error rate (5.1%) than smaller ones (5.6%).

A comparison of the reaction time means in Table 3 shows that the magnitude of the category size effect for Set 1 triples was considerably larger than the reverse effect for Set 2 triples (147 versus 57 milliseconds). This asymmetry appears, at first blush, to offer some evidence for the role of category size. But another reason for this asymmetry is suggested by inspection of the actual production frequencies of the category terms in the Loftus and Scheff norms. The difference in production frequency (similarity) between an instance and its two categories was also smaller for Set 2 than for Set 1 triples, and this could explain why the reaction time effect for Set 2 triples was less than that for Set 1 triples. To obtain a more accurate view of the role of similarity on categorization latencies, we performed an analysis of covariance, similar to the analysis of variance just reported, with the Loftus and Scheff production frequency as a covariate. In addition, since there was a possibility of the confounding of word frequencies with similarity, we used the frequencies of the superordinate terms (Kučera & Francis, 1967) as a second covariate. The results of this analysis were unambiguous. When both covariates were entered into the analysis, production frequency accounted for a significant portion of the variance ($F = 22.56$, $df = 1/22$, $p < .01$), while word frequency was insignificant ($F < 1$, $df = 1/22$). Further, when production frequency was used as the only covariate, the critical interaction between type of triple and set size was reduced to below conventional significance levels ($F = 3.07$, $df = 1/23$, $.05 < p < .10$), while when word frequency was used as the single covariate, the size of the critical interaction was relatively unchanged ($F = 10.09$, $df = 1/23$, $p < .01$). Thus production frequency,

rather than word frequency, accounts for the obtained reaction time differences.[7] Note, too, that when either factor serves as covariate, the effects of overall category size remain insignificant ($F < 1$, $df = 1/23$, when production frequency covaries; $F = 2.26$, $df = 1/23$, $p > .10$, when word frequency covaries). All things considered, the results of this experiment on category size are consistent with the feature comparison model and problematic for the approaches taken by Collins and Quillian (1969; 1972a), Meyer (1970), and Landauer and Meyer (1972).

There is one further issue to be considered in regard to this experiment. We have assumed that the production frequency associated with each *true* item measures the semantic relatedness of that instance–category pair, and it would be helpful to have some independent evidence to support this assumption. The simplest form this evidence could take would be a demonstration that, for the items in question, production frequencies correlate with direct ratings of semantic relatedness. In order to measure the correlation between these two variables, we had the 52 *true* items rated for their semantic relatedness by a new group of 100 Stanford undergraduates. Fifty of these subjects rated the relatedness of the 26 statements containing small categories, while the other 50 subjects performed the comparable ratings for the statements containing large categories. (Our reason for having different subjects rate small- and large-category statements was that pilot data indicated that when the same subject rated all 52 statements, he tended to use category size as a cue in making his relatedness decision). The resulting mean ratings showed the same pattern as the mean production frequencies. Specifically, for Set 1 triples, the mean rating for the small-category statements (6.05) was greater than that for large-category statements (5.22), while for Set 2 triples, small-category statements were rated on the average as less similar than their large-category counterparts (6.36 versus 6.58). To assess further the relation between these ratings and the corresponding production frequencies, we proceeded as follows. We first took the difference, in ratings and in production frequencies, between statement pairs of the form *An S is a* P_L and *An S is a* P_S, where P_L is the larger category of S, and P_S is the smaller category. (The purpose of using this difference score was to minimize effects due solely to the S term.) We then correlated these difference scores for ratings with those for production frequencies. The resulting correlation was .47 ($df = 24$, $p < .05$). This provides some support for our assumption that production frequency reflects semantic similarity. But the correlation is quite small when compared to comparable findings; for example, Rips et al. (1973) report a correlation of .85 between ratings and production frequencies. Also, subsequent analysis of the present data indicated that *true* reaction times correlated higher with the production frequencies than with the ratings. Thus while it appears that both production frequencies and ratings of relatedness measure semantic relatedness, we do not understand fully what determines the amount of agreement between these measures or what determines which will be the more sensitive predictor of reaction time.

Additional Category Size Effects

While the previous findings provide evidence for the feature comparison model, there are two results in the literature on category size that pose potential problems for the model. One such troublesome finding concerns *true* reaction time (Wilkins, 1971) and the other, *false* reaction time (Landauer & Meyer, 1972).

We deal first with the problem regarding *true* reaction time. According to the feature comparison model, category size should increase *true* times only when increasing size also decreases overall similarity. It follows that, if one increases category size while

[7] There are indications, though, that frequency can affect *false* reaction time. Both Smith (1967) and Landauer (personal communication, 1972) have found that the more frequent a nontarget instance, the more rapidly a subject can correctly decide that it is not a member of the target category. It is possible, however, that this frequency effect may be partly mediated by a relatedness effect.

holding overall similarity constant, then there should be no increase in *true* reaction time. Rips et al. (1973) demonstrated this statistically by showing that there is no significant category size effect in an analysis of variance if similarity ratings are used as a covarying factor. However, Wilkins (1971) has presented data which suggest a standard category size effect even when overall similarity is held constant. Wilkins varied category size not by using nested categories but rather on the basis of norms which specify how many instances subjects produce for a particular category. To gain some control over overall similarity, Wilkins used as positive instances either an instance that was the most frequently produced in the norms (high similarity), or one that was rarely produced (low similarity). The result of central concern was that *true* reaction time increased with category size for both high and low similarity instances. Thus there is a category size effect even when overall similarity is apparently held constant.

But the Wilkins study is open to three criticisms. First, Wilkins' method of varying category size is suspect as there is no assurance that his larger categories always contained more instances than his smaller ones. Indeed, as Landauer and Meyer (1972) note, if *month* and *animal* are used as categories in a production situation, subjects may actually produce more instances of *month* than of *animal* (Freedman & Loftus, 1971). Second, Wilkins' method of equating similarity across categories is equally suspect. The fact that instance I_1 is the most frequent response to category C_1, while I_2 is the most frequent response to category C_2 does not insure that the overall similarity between the I_1–C_1 pair equals that between the I_2–C_2 pair. That is, Wilkins attempted to equate similarity across categories by using ordinal rather than interval measures, and hence, any obtained category size effect could still be due to similarity.[8] Our final criticism is simply that the magnitude of Wilkins' category size effect is but 17 milliseconds, and given the estimates of item variability in semantic memory studies considered by H. Clark (1973), there is some doubt that this result would generalize to a new set of items.[9]

Consider now findings concerning category size and *false* reaction time. As mentioned earlier, *false* reaction time has been shown to increase with category size (e.g., Landauer & Freedman, 1968; Meyer, 1970), and we have not yet offered an explanation of this finding in terms of the feature comparison model. The general explanation, which was first suggested by Collins and Quillian (1970), is as follows: For a fixed set of nontarget instances, as one increases the size of the target category there may also be a concommitant increase in the similarity or relatedness of the test instance and category. To illustrate, consider the case where the instance *vegetable* has to be verified against *dog* or *animal*. With the incease in category size from *dog to animal*, there is a presumed increase in the relatedness from *vegetable–dog* to *vegetable–animal*. Since *false* reaction time is known to increase with relatedness, it follows that an increase in category size can lead to an increase in *false* times.

[8] To make this argument more concrete, consider something of a limiting case. Suppose that a fixed set of instances, $I_1 \ldots I_n$, are rated for semantic similarity with respect to two superordinate categories C_1 and C_2. Suppose further that the results are such that the similarity rankings of instances for C_1 are identical to those of C_2, but the actual values are such that each instance is more similar to C_1 than to C_2. That is, the similarity of any instance to C_1 exceeds the similarity of that instance to C_2 by a constant amount. In this case, we would certainly expect faster reaction times to C_1 than to C_2. For even though similarity has been controlled at the level of ranks, it is still varying between categories when the interval properties of the ratings are considered. Something approaching this limiting case has been found in our unpublished data. When mammal instances are rated for similarity with respect to the categories *mammal* and *animal*, the similarity rankings for the two categories are positively correlated. But the ratings for the animal category are consistently higher than those for the mammal category, and the *true* reaction times are also consistently faster for the *animal* than *mammal* category.

[9] Although our last two criticisms apply to his published paper, Wilkins (personal communication, March, 1974) indicated that he has performed new analyses. These analyses support his initial interpretation.

Again, we are led to the prediction that if one increases category size but holds similarity (relatedness) constant, there should be no increase in reaction time. Yet, Landauer and Meyer (1972) have attempted to hold relatedness constant and have still shown a category size effect. In this experiment, category size was varied by means of nested triples. Relatedness was varied by having subjects rank the relatedness of each nontarget instance to the category it was used with and then averaging the ranks over subjects to dichotomize the items into related and unrelated sets of items. The critical findings were that *false* reaction time increased with category size for both related and unrelated items, while relatedness per se had no effect. The failure to obtain a relatedness effect on *false* reaction time is suspect, in view of the findings reported earlier (Basic Predictions and Results), and suggests that most of the *false* instance–category pairings employed may have been unrelated. This suggestion is supported by the fact that the rankings of relatedness for different subjects had a mean concordance of only .27. A concordance of .27 implies that the mean Spearman rank-order correlation coefficient between rankings for a given category was around .20 (see, e.g., Winer, 1971, p. 303), which in turn suggests that there was a rather small amount of shared variance between the rankings of different subjects (R. J. Sternberg, personal communication, 1973).

Even assuming that the *false* items were all unrelated, we are still left with the problem of why Landauer and Meyer obtained a category size effect that cannot be attributed to relatedness. One possibility is that the effect is due to some extraneous aspects of some of the items. Specifically, the magnitude of the effect is relatively small (32 milliseconds), and there is no statistical indication of whether this effect generalizes to other items (H. Clark, 1973). Another possibility is that the effect reflects the subject's certainty about his decision rather than a search of semantic categories. That is, a subject may be less certain in deciding that an item is not a member of a large abstract category than in deciding that this same item is not a member of a smaller and more concrete category. This lack of certainty could arise from the abstractness of the predicate categories and could lead to a selective slowing of *false* reaction time, independent of any category search. (In terms of the feature comparison model, this selective slowing could be implemented by changing the setting of the low criterion, c_0.) Both of these possible explanations of the Landauer and Meyer result must be considered tenuous, however, since we can offer little independent support for either. Though the Landauer and Meyer (1972) result is some cause for concern, the feature comparison model still seems to offer a reasonable account of a wide variety of category size results.

A Quantitative Test of the Model

Quantitative Formulation

In the previous sections we have tried to demonstrate that the feature comparison model is sufficient to explain many of the qualitative findings that have been obtained in a variety of situations. But there is another kind of sufficiency that one may ask of a model. This is its ability to predict quantitative aspects of data within a set of conditions that clearly fall within the boundaries of the model, and this is the concern of the present section. Quantifying the present proposal in a straightforward matter by means of a generalization of the mathematical treatment provided by Atkinson and Juola (1973, 1974) for a similar process.

Let us begin by reiterating the basic processes involved in verifying an instance–category relation. The subject first retrieves the sets of semantic features for the instance and category and then determines the overall semantic similarity, x, of the category to the instance. The subject uses this x value to decide whether to execute a fast *true* response ($x > c_1$), or a fast *false* response ($x < c_0$), or to go on to a second comparison stage ($c_0 \leq x \leq c_1$) that considers defining features of the instance and category. Figure 4 contains a schematic representation of this process for the illustrative case where a subject determines whether test instances are

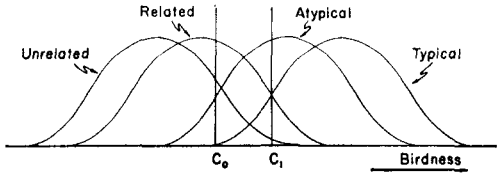

FIGURE 4. Hypothetical distributions of overall similarity values when target category is *bird*.

members of the category *bird*. Only two target instances are shown, one typical and one atypical, and only two nontargets, one related to the target category and one unrelated. Each of the four functions is a distribution of x values. All distributions are plotted on a continuum we have called "birdness" because the first comparison process determines overall semantic similarity with respect to that target category. The leftmost distribution represents the nontarget that is unrelated to *bird*, the next distribution is for the related nontarget, the third distribution pertains to the atypical bird instance, and the rightmost distribution is for the typical instance. The ordinal placement of these distributions along the birdness continuum reflects only our previous assumptions about shared defining and characteristic features. It further follows from these assumptions that a greater portion of the typical than the atypical distribution will exceed c_1 and that a greater portion of the unrelated than the related distribution will be less than c_0. These two inequalities give rise to the typicality effect for correct *true* reaction time and the relatedness effect for correct *false* reaction time, respectively.

Quantifying the model amounts to deriving expressions for the probability of each type of error, as well as expressions for correct *true* and *false* reaction time. Starting with the derivations for error probabilities, we first assume that the density function for each type of item is a normal distribution with unit variance, reflecting the probability that the first stage will yield a particular value, x. We further assume that the mean of each distribution, μ_i, is related by some monotone transformation to subjects' ratings of the semantic similarity between test instance i and the target category. Finally, we assume that all errors reflect first-stage processing; that is, a subject will erroneously respond *false* to a target instance only if the x value associated with that instance is below c_0 and will erroneously respond *true* to a nontarget only if the associated x value is greater than c_1. Given these assumptions, the probability of an erroneous *false* response to target instance i can be expressed as follows:

$$P(\textit{false}\,|\,\text{target i}) = \int_{-\infty}^{c_0} \phi(x - \mu_i)\,dx = \Phi(c_0 - \mu_i). \quad [1]$$

Here ϕ is the normal density function with unit variance and mean zero, and Φ is the corresponding distribution function. Similarly, the probability of an erroneous *true* response to a nontarget item j is given by the integral of $\phi\ (x - \mu_j)$ from c_1 to ∞,

$$P(\textit{true}\,|\,\text{nontarget j}) = \int_{c_1}^{\infty} \phi(x - \mu_j)\,dx = 1 - \Phi(c_1 - \mu_j). \quad [2]$$

Thus, to predict obtained error rates, we need to consider estimates of the criteria and the relevant means.

Expressions for correct reaction times can be developed along similar lines. For each correct response, mean *true* reaction time is a function of a base time, t_T, equal to the time needed for first-stage processing and execution of a positive response, and a second interval, t_2, which reflects the duration of second-stage processing. This second stage will occur only on those errorless trials in which x falls between c_0 and c_1. Mean *true* reaction time (RT) to target i is therefore equal to t_T plus the value of t_2 weighted by the probability of occurrence of the second stage:

Mean Correct *true RT*

$$= t_T + t_2\left[\frac{\int_{c_0}^{c_1} \phi(x - \mu_i)dx}{\int_{c_0}^{\infty} \phi(x - \mu_i)dx}\right]$$

$$= t_T + t_2\left[\frac{\Phi(c_1 - \mu_i) - \Phi(c_0 - \mu_i)}{1 - \Phi(c_0 - \mu_i)}\right]. \quad [3]$$

Here, the expression in brackets represents the probability of executing the second stage, conditionalized on a correct response. The same assumptions apply for correct *false* reaction time, leading to the following equation for nontarget item j:

Mean Correct *false RT*

$$= t_F + t_2\left[\frac{\int_{c_0}^{c_1}\phi(x-\mu_j)dx}{\int_{-\infty}^{c_1}\phi(x-\mu_j)dx}\right]$$

$$= t_F + t_2\left[\frac{\Phi(c_1-\mu_j)-\Phi(c_0-\mu_j)}{\Phi(c_1-\mu_j)}\right] \quad [4]$$

In this case, t_F represents the time necessary for first-stage processing and execution of a negative response, and the expression in brackets again represents the probability of executing the second stage, conditionalized on a correct response.

So, to predict obtained reaction times we must consider the time parameters t_T, t_F, and t_2 in conjunction with the parameters already introduced to handle error rates. However, there is an additional consideration which suggests the need for another time parameter. Although the time needed for second-stage processing is represented by a single parameter in Equations 3 and 4, there is reason to believe that this value should change as a function of the category size of the target category. We have previously noted that the larger the target category, the fewer the number of defining features which need to be compared in the second stage. For this reason we expect that second-stage processing time for larger categories should be less than for smaller categories. We can distinguish, then, two values, $t_{2,S}$ and $t_{2,L}$, which represent second-stage processing time for small and large categories, respectively.[10]

[10] One could also argue that the time needed for second-stage processing is affected by whether the process eventuates in a positive or negative decision. This suggests that separate t_2 parameters should be estimated for *true* and *false* items. However, to achieve some parsimony, we assume that this effect of truth value will be negligible in the second stage and concern ourselves only with the two t_2 parameters mentioned in the text.

Within the framework of Equations 1 through 4, there are a number of ways in which predictions can be derived. One simple way is akin to the method used by Atkinson and Juola (1973). In this method we bypass the possibility of using semantic similarity ratings to estimate means and instead use the observed error rates to estimate reaction time by substituting these error rates on the left-hand side of Equations 1 and 2. This allows us to solve for the quantities $c_0 - \mu_1$ and $c_1 - \mu_j$. We need only estimate the difference $c_1 - c_0$ in order to fix the relative positions of c_0, c_1, μ_1, and μ_j. Since the absolute values of these parameters are irrelevant in terms of the four equations, one of these parameters may be set at an arbitrary value. If we set $c_0 = 0$, then estimation of c_1 for each target category, together with the four time parameters, t_T, t_F, $t_{2,S}$, and $t_{2,L}$ (assumed to be invariant across categories), is sufficient to predict mean reaction times on the basis of Equations 3 and 4. A second method of obtaining predictions from this model involves the use of subjects' ratings of the semantic similarity of the various instance–category pairs. By this method we must first obtain estimates of c_1 for each target category and estimates of the distribution means of the various instances within each block of trials. Error rates can be predicted in this way from Equations 1 and 2. These same estimates can then be used in conjunction with estimates of the four time parameters to predict mean reaction time. We will use both of these methods to predict results for the following experiment.

A Test of the Model: Experiment 2

To provide data to test the quantitative model, a straightforward instance–category verification experiment was conducted. We first selected four categories—*bird, insect, fruit,* and *vegetable*—and 12 target instances for each of these categories. In the task, each of the four categories served as a target category for a block of 24 trials, 12 of these trials containing the target instances and the other 12 containing nontargets. The 12 nontargets for a block were the instances of that category most related to the target.

TABLE 4

AVERAGE OBTAINED CORRECT TRUE REACTION TIME AND ERROR PERCENTAGE AS A FUNCTION OF TYPICALITY, SEPARATELY FOR EACH CATEGORY

Target category	Instance	High typicality		Medium typicality		Low typicality	
		Reaction time	% error	Reaction time	% error	Reaction time	% error
Bird	birds	505 (510)	4	510 (514)	5	535 (529)	8
Insect	insects	570 (562)	16	535 (562)	16	598 (580)	22
Fruit	fruits	518 (518)	2	583 (586)	14	653 (650)	38
Vegetable	vegetables	623 (627)	10	670 (681)	26	713 (694)	32
Animal	birds and insects	528 (557)	6	568 (558)	6	610 (586)	16
Plant	fruits and vegetables	553 (553)	12	546 (560)	14	564 (548)	10
Mean		550 (554)	8	569 (577)	15	612 (598)	21

Note. Predicted correct *true* reaction times are in parentheses.

That is, when *bird* (*insect*) was the target, the nontargets were the insect (bird) instances, while when *fruit* (*vegetable*) served as a target, the nontargets were the vegetable (fruit) instances. In addition, we included a block of 48 trials in which the target was *animal* and a comparable block where *plant* served as target category. For the *animal* block, the 24 target trials contained the bird and insect instances, while the nontargets consisted of the fruit and vegetable instances; for the *plant* block, this assignment of target and nontarget instances was reversed. Our reason for including these two target categories was to see if the model could handle category size effects. There were, then, six blocks in all, corresponding to the six possible target categories, and the order of the instances in each block was randomized. The order of the blocks was partially counterbalanced across subjects, and the assignment of responses (*true* versus *false*) was balanced over handedness across subjects. The target category was given verbally at the start of a block, and each test instance was presented visually to a subject who determined whether it belonged to the target category or not. The subjects were 30 Stanford undergraduates.

To assess the degree of typicality of the target instances and the relatedness of nontargets for each of the six categories, we used the independent ratings of 29 Stanford undergraduates. These subjects rated how good a member each target and nontarget instance was of the relevant target categories on a seven-point scale. These ratings provide the basis for our estimates of distribution means in our second method of deriving predictions.

The results of the experiment are summarized in Table 4, which presents mean reaction time and error rate for *true* items as a function of typicality, separately for each category. Here, we have used the ratings to partition the instances of each category into three levels of typicality (high, medium and low) where each level contained an equal number of instances. Inspection of Table 4 indicates that, with but two exceptions, there is a consistent typicality effect on *true* reaction time for all categories. Moreover, there was also a remarkably consistent decrease in error rate with typicality for each category. This correlation between typicality, *true* reaction time, and error rate is exactly what the model predicts (see Figure 4).

One other aspect of the *true* reaction time data is worth noting. Though *true* times consistently decreased with typicality for small and large categories, it is not the case that the most typical instances of the small categories were necessarily the most typical instances of the large categories. This can be seen in an analysis of variance of *true* reaction time results using a repeated measures design over category size. This an-

TABLE 5

AVERAGE OBTAINED CORRECT FALSE REACTION TIME AND ERROR PERCENTAGE AS A FUNCTION OF RELATEDNESS, SEPARATELY FOR EACH CATEGORY

Target category	Instance	High relatedness		Medium relatedness		Low relatedness	
		Reaction time	% error	Reaction time	% error	Reaction time	% error
Bird	insects	605 (600)	19	555 (574)	10	553 (563)	8
Insect	birds	620 (598)	16	565 (588)	12	573 (588)	12
Fruit	vegetables	658 (684)	36	655 (640)	19	643 (593)	8
Vegetables	fruits	780 (758)	52	700 (725)	29	680 (696)	18
Animal	fruits and vegetables	580 (604)	9	589 (621)	15	612 (622)	15
Plant	birds and insects	640 (597)	14	615 (602)	17	608 (592)	12
Mean		647 (640)	24	613 (625)	17	611 (609)	12

Note. Predicted correct *false* reaction times are in parentheses.

alysis reveals a significant overall effect of typicality ($F' = 9.83$, $df = 2/65$, $p < .01$) and a significant inverse category size effect ($F' = 8.48$, $df = 1/50$, $p < .01$). But in addition, the interaction of these factors is also significant ($F' = 6.49$, $df = 2/72$, $p < .01$), indicating that instances typical of one category may be atypical of that category's superordinate. This documents the argument mentioned earlier (p. 220) in connection with the difficulties that alternative models have in handling typicality effects.

Table 5 presents mean reaction time and error rate for *false* items as a function of relatedness, separately for each category. Again, the instances used with each target category were partitioned into three levels (high, medium, and low relatedness), and each level included an equal number of instances. In general, both *false* reaction time and error rate tended to increase with relatedness, and again this correlation is predicted by the model. However, it should be noted that the results for the case in which *animal* served as target category (and fruit and vegetable instances served as nontargets) were inconsistent with the general pattern of results, and this inconsistency remains an anomaly.[11]

[11] There is one other qualitative result in the data in Tables 4 and 5 that deserves some comment. This result, was pointed out to us by A. Glass and K. Holyoak (personal communication, 1973), and it is convenient to describe it first for *true* reaction time. The *true* reaction time to any item tended to decrease when that item was unrelated to the category of the nontargets. As an example, consider the case where the target category was *bird*. *True* reaction time decreased as the relatedness between the target instance and the category *insect* decreased, and this effect was independent of the typicality of the target. Thus the *true* reaction time to any item was determined both by how typical that item was of the target category and how atypical the item was of the contrasting category. A comparable situation holds for *false* reaction time (though the effect in question is weaker). Thus the *false* reaction time to any item decreased with the typicality of that item to the contrasting category as well as with the atypicality of that item to the target category. While this result is an interesting one, it may be specific to the procedure used here (i.e., all nontargets were drawn from the same category). In any event, the result in question is not incompatible with the feature comparison model, but rather suggests, for example, that different subjects may have used different strategies with some subjects comparing the test item to the target category and others comparing the test item to the contrasting category.

Consider now the model's ability to predict these data quantitatively. Estimates of these latencies, using the first method described above, are shown in parentheses in Tables 4 and 5. The product moment correlation between predicted and observed reaction time is .945 ($df = 24$, $p < .01$), and the root mean square deviation (RMSD) for this set of data equals 28.9 milliseconds. Both of these statistics suggest that the fit

of the model is reasonable, but it is also of some interest to determine goodness of fit by a very stringent test. Such a test can be obtained by defining the following statistic for each of the predicted means:

$$T_i = \frac{\bar{X}_i - \hat{\bar{X}}_i}{\frac{s_i}{\sqrt{n_i - 1}}}. \qquad [5]$$

Here $\bar{X}$ is the obtained mean reaction time, $\hat{\bar{X}}$ the predicted mean reaction time, n_i the number of observations contributing to the obtained mean, and s_i the standard deviation of these observations. The sum $\sum_{i=1}^{k} T_i^2$ is distributed as chi-square with degrees of freedom equal to k minus the number of parameters used in fitting the k means. For the model considered here, the value of this statistic is 89.69, when all 36 means in Tables 4 and 5 are taken into account (with $df = 26$, $p < .005$). This, of course, is a disappointingly poor fit, but the major source of lack of fit is again the case in which vegetables and fruits appeared as nontargets for the category *animal*. Excluding the anamolous case yields more reasonable chi-square values: as examples, when only the smaller categories in Tables 4 and 5 are considered ($\chi^2 = 25.66$, $df = 15$, $.025 < p < .05$) and when only the 18 mean *true* reaction times in Table 4 are considered ($\chi^2 = 17.34$, $df = 9$, $.025 < p < .05$).

The time parameters estimated by this method of fitting the model are $t_T = 477$ milliseconds, $t_F = 514$ milliseconds, $t_{2,S} = 280$ milliseconds, and $t_{2,L} = 161$ milliseconds. In all cases, parameters were obtained by a computer search using a nonlinear least-squares technique. Note that, as predicted, second-stage processing time is less for the larger than the smaller target categories. In fact, these values for the second stage accord quite well with values obtained by Meyer (1970) using a very different set of assumptions. In obtaining all of these estimates we have assumed that errors are always the result of first-stage processing. Some evidence concerning this assumption can be gained from inspection of Table 6, which

TABLE 6

AVERAGE CORRECT REACTION TIME AND AVERAGE ERROR REACTION TIME (IN MILLISECONDS) FOR TARGET INSTANCES AND NONTARGET INSTANCES, SEPARATELY FOR EACH CATEGORY

Target category	Instance	Correct reaction time	Error reaction time
	Target instances		
Bird	birds	517	429
Insect	insects	568	562
Fruit	fruits	585	538
Vegetable	vegetables	669	701
Animal	birds and insects	569	502
Plant	fruits and vegetables	554	539
Mean		577	545
	Nontarget instances		
Bird	insects	571	470
Insect	birds	586	466
Fruit	vegetables	652	598
Vegetable	fruits	720	583
Animal	fruits and vegetables	594	444
Plant	birds and insects	621	483
Mean		624	507

presents a comparison between average correct and average error reaction time, separately for each category. If errors are based only on the first stage, then we would expect that, for each category, error reaction time will be less than or equal to the corresponding correct reaction time. With one exception (vegetable target items), this is the case. Furthermore, since the parameter t_F represents the time needed for first-stage processing and execution of a negative response, t_F should equal error times for target items; similarly, t_T should be approximately equal to error times for nontargets. In both cases, mean error reaction time is only 30 milliseconds longer than the predicted values (545 versus 514 and 507 versus 477). This suggests that second-stage processing may be reasonably error free, as assumed, at least for this type of categorization task.

In our second method of deriving predictions, we predict error rates by first using subjects' typicality ratings to determine

the means of the density functions and then combining these means with estimates of the criteria as specified by Equations 1 and 2. Since our preliminary attempts to fit this error data indicated that the relation between the ratings and the distribution means was nearly linear, these means were estimated by equations of the form

$$\mu_{i,C} = \alpha \bar{y}_{i,C} + \beta. \qquad [6]$$

Here $\bar{y}_{i,C}$ represents the mean typicality (or relatedness) rating of the instances of Group i (recall the instances are grouped in Tables 4 and 5) to Category C, and $\mu_{i,C}$ is the corresponding distribution mean. An important question arises as to whether the scaling parameters, α and β, should depend on the target category. In fact, the accuracy of the predictions improves considerably if α and β are estimated anew for each target category. In this case, the correlation between predicted and observed error rates is .934 ($df = 16$, $p < .01$) and the corresponding *RMSD* equals 5.1%. On the other hand, with α and β constant over all six categories, the correlation drops to .637 ($df = 26$, $p < .01$) with $RMSD = 11.7\%$. This indicates that typicality ratings are fairly sensitive predictors of error rates within a given category but that either ratings or errors are affected by properties of the target category which are independent of typicality.

To continue with our predictions from this method, we note that the positions of the distribution means and criteria have been determined by fitting our model to the error rate data, and thus reaction time predictions can be obtained by estimating the values of the four time parameters. These reaction time predictions, however, are considerably less accurate than those obtained by the first method we described. Even with separate values of α and β for each category, the correlation between predicted and obtained reaction time is only .690 ($df = 14$, $p < .01$) with $RMSD = 55.8$ milliseconds. This is to be compared with .945 and $RMSD = 28.9$ milliseconds obtained by the first method. This difference in predictive accuracy is attributable to the way in which the values of the upper criterion, c_1, are estimated. Specifically, in the first method, c_1 and the four time parameters are estimated simultaneously to determine reaction time predictions, while in the second method, c_1 must be estimated first in obtaining predicted error rates, leaving only the four time parameters to be estimated in a second step. For this reason, the second method produces relatively accurate estimates of error rates but does considerably less well in predicting reaction time.

In summary, using the formal version of our feature comparison model inherent in Equations 1 through 4, we can provide a relatively adequate quantitative account of typicality, relatedness, and category size effects in a semantic categorization task. However, our mathematics reflects only the more general aspects of the feature comparison model we have described in the previous sections. We consider this loss of theoretical detail to be a serious omission. In particular, in earlier sections we proposed a conception of second-stage processing which included the assumptions of separating out the more defining features of the category and then comparing them to the features of the test instance; but, in our mathematical treatment all of this processing was reflected by a single time parameter for a given size category. Had we tried to give these theoretical aspects some consideration in the mathematical formulation, we would have further complicated what is already a rather complex formulation. All things considered, the present results simply show that a straightforward formalization of the general aspects of our model is sufficient to account for many quantitative aspects of semantic categorization. It is, if you will, a demonstration of theoretical sufficiency. We hold that this is not a trivial demonstration, as witnessed by the lack of comparable predictions from other models of semantic memory.

SUMMARY AND EXTENSIONS

We began by assuming that the meaning of any lexical term includes characteristic as well as defining features and that superordinate terms contain fewer defining features than do their subordinate counter-

parts. This proposal is based, in part, on linguistic analyses, in particular Lakoff's (1972) analysis of hedges. This proposal was then shown to be consistent with typicality ratings, multidimensional space representations of these ratings, and acceptability judgments of sentence substitutions. To explain how subjects verify instance–category statements, we combined these structural assumptions with a two-stage processing model. In this model, the first stage assesses the overall semantic similarity (i.e., both defining and characteristic features) between the test instance and the target category, while the second stage determines the agreement between defining features of the category and those of the instance. This feature comparison model, which can yield either fast responses based on only single-stage processing or slower responses based on dual-stage processing, was shown to be consistent with several findings: the typicality effect on *true* reaction time, the conjoint frequency effect on *true* times (Wilkins, 1971), the relatedness effect on *false* times, the effect of set relation on *false* times (Meyer, 1970), and category size effects on both *true* and *false* times. With regard to the latter, we argued that under some circumstances the feature comparison model predicts reverse category size effects, and such reverse effects were demonstrated in a new experiment. Finally we developed a mathematical version of the feature comparison model and showed that it provided a quantitative account of some new experimental data.

So much for what has been done. It is our contention that the feature comparison model can be extended in a number of important ways. First, the model can clearly be applied to the results obtained in semantic memory tasks which involve a comparison of word meanings (Schaeffer & Wallace, 1969; 1970). A second application of our model concerns the verification of existentially quantified statements, for example, *Some women are writers,* which have been studied by Meyer (1970). The statements that we have dealt with in this article have all been true when universally quantified, but the basics of the feature comparison model are also applicable to statements that are true only when existentially quantified. A third avenue of extension is to statements that assert something about a property, for example, definitionally true property statements like *A canary is yellow* (e. g., Collins & Quillian, 1969). These extensions, and related matters, are treated in depth in Smith et al. (1974), where the feature comparison model is used as a vehicle for exploring numerous issues in psychological studies of semantics.

In addition to these specific applications of the feature comparison model, the structural assumptions of the model are appropriate to the study of inductive and analogical thought as well as to logical reasoning (see Smith et al., 1974). For example, Rips et al. (1973) demonstrated that performance in an analogies task could be predicted on the basis of the relations between the characteristic features of the terms in the analogy. In this case, however, the processing assumptions were based on Rumelhart and Abrahamson's (1973) theory rather than on the processing model used in the present article. But the fact that different processing assumptions are needed for qualitatively different tasks should not disturb us. Rather, what seems to be more important is that an approach to semantic memory based on semantic features may be capable of encompassing a truly wide variety of semantic phenomena that are of interest to both psychologists and linguists.

REFERENCES

ATKINSON, R. C., & JUOLA, J. F. Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance.* Vol. 6, New York: Academic Press, 1973.

ATKINSON, R. C., & JUOLA, J. F. Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology.* San Francisco: Freeman, 1974, in press.

BATTIG, W. F., & MONTAGUE, W. E. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monograph,* 1969, **80**, (3, Pt. 2).

BELL, A., & Quillian, M. R. Capturing concepts in a semantic net. In E. L. Jacks (Ed.),

Associative information techniques. New York: Elsevier, 1971.

BIERWISCH, M. Some semantic universals of German adjectivals. *Foundations of Language,* 1967, **3**, 1–36.

BIERWISCH, M. On classifying semantic features. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology.* Cambridge, England: University Press, 1971.

BOLINGER, D. The atomization of meaning. *Language,* 1965, **41**, 555–573.

CARROLL, J. D., & CHANG, J. J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika,* 1970, **36**, 283–319.

CLARK, E. V. What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language.* New York: Academic Press, 1973.

CLARK, H. H. Linguistic processes in deductive reasoning. *Psychological Review,* 1969, **76**, 387–404.

CLARK, H. H. Word associations and linguistic theory. In J. Lyons (Ed.), *New horizons in linguistics.* Baltimore: Penguin, 1970.

CLARK, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior,* 1973, **12**, 335–359.

COHEN, B. H., BOUSFIELD, W. A., & WHITMARSH, G. A. Cultural norms for verbal items in 43 categories. (Office of Naval Research Tech. Rep. 22, Nonr 631(00) Storrs: University of Connecticut, 1957.

COLLINS, A. M., & QUILLIAN, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 240–248.

COLLINS, A. M., & QUILLIAN, M. R. Does category size affect categorization time? *Journal of Verbal Learning and Verbal Behavior,* 1970, **9**, 432–438.

COLLINS, A. M., & QUILLIAN, M. R. Experiments on semantic memory and language comprehension. In L. W. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley, 1972. (a)

COLLINS, A. M., & QUILLIAN, M. R. How to make a language user. In E. Tulving & W. Donaldson (Eds.), *Organization of memory.* New York: Academic Press, 1972. (b)

CROSS, D. V. Metric properties of multidimensional stimulus generalization. In I. Mostofsky (Ed.), *Stimulus generalization.* Stanford: Stanford University Press, 1965.

EGETH, H. E. Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics,* 1966, **1**, 245–252.

FREEDMAN, J. L., & LOFTUS, E. F. Retrieval of words from long-term memory. *Journal of Verbal Learning and Verbal Behavior,* 1971, **10**, 107–115.

GARNER, W. R., & FELFOLDY, G. L. Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology,* 1970, **1**, 225–241.

GLASS, A. L., HOLYOAK, K., & O'DELL, C. Production frequency and the verification of quantified statements. *Journal of Verbal Learning and Verbal Behavior,* 1974, in press.

HENLEY, N. M. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior,* 1969, **8**, 176–184.

KATZ, J. J., & FODOR, J. A. The structure of a semantic theory. *Language,* 1963, **39**, 170–210.

KRUSKAL, J. B. Multidimensionl scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* 1964, **29**, 1–27.

KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English.* Providence: Brown University Press, 1967.

LAKOFF, G. 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, 1972.

LANDAUER, T. K., & FREEDMAN, J. L. Information retrieval from long-term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior,* 1968, **7**, 291–295.

LANDAUER, T. K., & MEYER, D. E. Category size and semantic memory retrieval. *Journal of Verbal Learning and Verbal Behavior,* 1972, **11**, 539–549.

LOFTUS, E. F. Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology,* 1973, **97**, 70–74.

LOFTUS, E. F., & SCHEFF, R. W. Categorization norms for 50 representative instances. *Journal of Experimental Psychology,* 1971, **91**, 355–364.

MEYER, D. E. On the representation and retrieval of stored semantic information. *Cognitive Psychology,* 1970, **1**, 242–299.

MILLER, G. A. A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology,* 1969, **6**, 169–191.

MILLER, G. A. English verbs of motion: A case study in semantic and lexical memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory.* Washington, D. C.; Winston, 1972.

RIPS, L. J., SHOBEN, E. J., & SMITH, E. E., Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior,* 1973, **12**, 1–20.

ROSCH, E. On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and acquisition of language.* New York: Academic Press, 1973.

ROSCH, E. Universals and cultural specifics in human categorization. In R. Breslin, W. Loner, & S. Bochner (Eds.), *Cross-cultural perspectives.* London: Sage Press, 1974, in press.

RUMELHART, D. E., & ABRAHAMSON, A. A. Toward a theory of analogical reasoning. *Cognitive Psychology,* 1973, **5**, 1–28.

RUMELHART, D. E., LINDSAY, P. H., & NORMAN, D. A. A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory,* New York: Academic Press, 1972.

SAPIR, E. Grading: A study in semantics. *Philosophy of science,* 1944, **11**, 93–116.

SCHAEFFER, B., & WALLACE, R. Semantic similarity and the comparison of word meanings. *Journal of Experimental Psychology,* 1969, **82**, 343–346.

SCHAEFFER, B., & WALLACE, R. The comparison of word meanings. *Journal of Experimental Psychology,* 1970, **86**, 144–152.

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika,* 1962, **27**, 125–140. (a)

SHEPARD , R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, II. *Psychometrika,* 1962, **27**, 219–246. (b)

SHEPARD, R. N. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology,* 1964, **1**, 54–87.

SMITH, E. E. Effects of familiarity on stimulus recognition and categorization. *Journal of Experimental Psychology,* 1967, **74**, 324–332.

SMITH, E. E., RIPS, L. J., & SHOBEN, E. J. Semantic memory and psychological semantics. In G. H. Bower (Ed.), *The psychology of learning and motivation.* Vol. 8. New York: Academic Press, 1974, in press.

TOWNSEND, J. T. A note on the identifiability of parallel and serial processes. *Perception & Psychophysics,* 1971, **10**, 161–163.

TULVING, E. Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory.* New York: Academic Press, 1972.

WILKINS, A. T. Conjoint frequency, category size, and categorization time. *Journal of Verbal Learning and Verbal Behavior,* 1971, **10**, 382–385.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1971.

WITTGENSTEIN, L. *Philosophical investigations.* (Trans. by G. E. M. Anscombe) Oxford, England: Blackwell, 1953.

(Received June 7, 1973; revision received December 18, 1973)