JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 15, 103–117 (1976)

Subjective Structures and Operations in Semantic Memory

ALFONSO CARAMAZZA, HARRY HERSH, AND WARREN S. TORGERSON

The Johns Hopkins University

Relations of semantic distance to response latencies in similarity judgments, to reaction times in a same-different classification task, and to proximity of recall in a free recall task were investigated. A multidimensional analysis of the similarity judgments was used to determine semantic structure and distances for pairs of animal words chosen from three classes: *mammals*, *birds*, and *fish*. A four-dimensional structure revealed both the class and quantitative contributions to the similarity ratings. Semantic distances were directly related to response latencies for words from the same class and inversely related for words from different classes. Distances were also related to reaction times for "same" judgments in the classification task, but not for "different" judgments. Independent ratings of typicality did not improve the relationship. Semantic distance was also a good predictor of proximity of recall for mammals, less so for birds, and not at all for fish.

A number of recent studies have been directed at explicating the structure of semantic memory and the nature of the operations that take place when retrieving information from long-term memory (Collins & Quillian, 1972; Meyer, 1970; Schaeffer & Wallace, 1970; Smith, Shoben, & Rips, 1974). The dependent variable used in these recent studies has usually been reaction time (*RT*) and the tasks that have been employed are: (*a*) to verify statements about class membership (Collins & Quillian, 1972; Rips, Shoben, & Smith, 1973; Schaeffer & Wallace, 1969, 1970; Smith et al., 1974); (*b*) to verify universally (*all*, *no*) and existentially (*some*) quantified statements (Meyer, 1970; Glass, Holyoak & O'Dell, 1974; Glass & Holyoak, 1974); and (*c*) to determine whether two strings of letters are both words (Meyer & Schvaneveldt, 1971).

A dominant factor that has emerged from all these studies is that the time to make a correct response is determined, in part, by "semantic distance" between instance and category in the first two tasks and between words in the third task. In this paper we examine the effects of semantic distance in a structured subset of the lexicon, selected instances of the three categories *birds*, *fish*, and *mammals*, on performance of several different tasks.

There are many ways in which the semantic distance between words can be estimated. One common way used in many of the studies cited earlier is to define semantic distance using common-sense relations holding between words. Thus, for example, Schaeffer and Wallace (1970) simply asserted that *canary* and *bird* are semantically more similar or close than *canary* and *animal* since the members of the first pair share more meaning components (e.g., feathered, fly, are winged, lay eggs) than the second pair. A more direct way of estimating the semantic distance between words is to have subjects themselves judge the semantic distance between words. This approach, together with the appropriate analytic techniques (e.g., clustering and scaling), has been used with considerable success in revealing interesting properties of the subjective lexicon (Miller, 1967, 1969,

The research reported in this study was supported by PHS Grant 5 TO 1 GMO2148 to The Johns Hopkins University. The authors are grateful to Bert Green, Jr., Ellen Grober, Jack Yates and Edgar Zurif for comments on an earlier version of the paper and to Sam Miller for his assistance in data collection.

1972; Henley, 1969; Fillenbaum & Rapaport, 1971). What is most interesting, however, is that knowledge of these subjective structures and the associated distances between words can be used to predict various kinds of language performance (Rips et al., 1973; Rumelhart & Abrahamson, 1973; Smith et al., 1974; Zurif, Caramazza, Myerson, & Galvin, 1974). Several of these studies are relevant to the research we are reporting and are briefly reviewed here.

Rumelhart and Abrahamson (1973) have used an experimentally derived structure to predict performance on analogical reasoning problems. They hypothesized that subjects can operate upon elements within a multidimensional representation using the Euclidean distances between elements as directed vectors. Employing the mammal configuration obtained by Henley (1969) and Luce's choice model (Luce, 1959), the probability of each analogy completion alternative was successfully predicted. In another experiment in the same study, fictitious animals were created by selecting arbitrary points in the three-dimensional configuration. Through a paired-associate paradigm incorporating the fictitious animals in incomplete analogies, subjects were able to form conceptualizations of these nonexistent mammals. When subjects were asked to describe the animals, descriptions referring to the animals' size and degree of ferocity corresponded to the relative values along the dimensions of the multidimensional space.

A second set of related studies have been reported by Smith and his co-workers (Smith et al., 1974; Rips et al., 1973). In these studies they first obtained separate multidimensional representations for each of two separate sets of animal terms (*birds* and *mammals*). The two solutions were each interpretable in two dimensional spaces where the dimensions could be interpreted as *size* and *ferrocity*. The Euclidean distances between points in the respective spaces (e.g., *hawk–cardinal, lion–mammal*) were then used to predict *RT* in several categorization experiments. Generally, it was found that semantic distances accounted for a statistically significant proportion of the variation in *RT*. In a separate experiment they attempted to extend the findings of Rumelhart and Abrahamson (1973) on analogical reasoning to include operations across categories (*mammals* and *birds*).[1] It seems, then, that semantic distance is an important variable in predicting performance in tasks that depend on retrieval of information from semantic memory.

These previous studies (with the possible exception of Rips et al., 1973, see Note 1, however) have constrained the areas under study to single, well-defined categories (e.g., *mammals* and *birds*). Obviously, people have the ability to deal with more than one category at a time, and can make comparisons and generally operate both within and across

[1] There are a number of flaws in this study that weaken any conclusion that might otherwise have been drawn from it. First, the rescaling of a subset of the points using the data obtained from judgments in a larger context is not a legitimate procedure. Arnold (1971) has shown that the final spatial solution one obtains in multidimensional scaling is context sensitive. Thus, there is no guarantee that independent judgments and scaling of the six animal terms used in this experiment would have given results similar to those obtained in the Rips et al. study. A second objection concerns their use of separate spaces for each animal category (*birds* and *mammals*). Again there is no assurance that the dimensions and relative locations of each item in each subspace obtained by scaling the two categories separately would have been the same had they scaled the twelve terms together. In addition there is the problem of poor fits: A correlation of .77 reported in this study shows substantial distortion was involved in fitting the points to the space. Finally, and of most importance, there is the problem of artificiality that could be raised against this latter study. Unlike the Rumelhart and Abrahamson experiment where all the terms in the space were used in the analogical task, Rips et al. selected for investigation pairs of terms that stood in symmetrical relation to each other across the two subspaces. This latter manipulation has the effect of making impossible a test of the Rumelhart and Abrahamson model in that there were no points close enough to the correct choice to test the vector hypothesis.

well-known categories. While the multidimensional model could reasonably represent the organization of semantic memory for some within-category tasks, could the model be extended to a more general semantic field incorporating several related classes? And if such a representation were obtainable, would it reflect performance on tasks generalized to the enlarged domain?

The following series of experiments were undertaken to investigate subjective organization both within and across related categories, and to examine subjects' ability to operate upon elements within this structure. The first experiment sought to derive a spatial structure from a multidimensional analysis of similarity judgments of a set of animal terms that could be subdivided into three distinct phylogenetic classes, *birds*, *fish*, and *mammals*. Judgment latencies were also obtained as a measure of the difficulty encountered in the judgment process. Experiment II employed a categorization procedure similar in many respects to that reported by Rips et al., except that three categories were tested simultaneously. In the third experiment typicality ratings were obtained in order to test predictions derived from a recent model proposed by Smith et al. (1974). Experiment IV made use of a free-recall task to test further the effects of semantic distance on a noncategorization task.

Experiment I: Similarity Ratings

Method

Stimuli. Thirty animal terms, 10 each of *fish*, *birds*, and *mammals* were selected from the Battig and Montague norms (1969). (Lacking *mammal* norms, 10 animals were chosen from the category *four-footed land animals*, a subjectively equivalent category.) The terms were arbitrarily chosen with the following constraints: (1) a term could not contain the category name (e.g., *bluebird*, catfish), (2) a term had to be sufficiently familiar to insure that subjects were familiar with the concept denoted by the term, and (3) terms were chosen to represent a comprehensive range of exemplars for the three categories under consideration. The list of animals, shown in Table 1, were roughly equated across categories for conjoint frequency (Battig & Montague, 1969).

TABLE 1

Categories with Ten Instances used in Experiment

Birds	Fish	Mammals
Canary	Barracuda	Bull
Crow	Flounder	Camel
Eagle	Herring	Elephant
Hawk	Mackerel	Horse
Owl	Marlin	Lion
Pelican	Minnow	Mouse
Pheasant	Pike	Rabbit
Robin	Salmon	Rhinoceros
Stork	Shark	Tiger
Vulture	Tuna	Wolf

Subjects. Fifteen students at The Johns Hopkins University (11 males) served as paid subjects.

Procedure. The experiment was computer controlled using a DEC PDP-11/20 and a Tektronics 4010 CRT terminal. The computer randomized the presentation of the 435 stimulus pairs (i.e., all combinations of 30 words) and also randomly assigned left–right locations of pair members on the screen so that temporal and spatial positions of the words were counterbalanced across pairs and subjects. The computer was also used to display instructions to the subjects.

The subjects were told that all possible pairs of 30 animal terms would be presented on the screen, one pair at a time. The subjects' task was to look at both animal terms and decide on a scale from 1 to 9 just how similar the two terms were to him. A rating of 1 indicated high similarity; 9 indicated high dissimilarity. They were told to answer as quickly as possible and not to worry about being consistent. Subjects were unaware that their judgment latencies were being recorded.

Since the response buttons were the nine numeral keys at the top of the terminal keyboard, it was felt that a subject's response latencies might be affected by the location of the keys. To obtain an index of the difference in latencies due to key position, a reaction time task was run on each subject at the end of the session. (This task was delayed until the end of the rating study because subjects were not informed that their speed of response was being recorded, and any suspicion that times were being recorded might bias the subjects' response strategies.) For this task 45 digits (five for each key position) were randomly flashed on the screen, one at a time. The subject simply had to type the key which corresponded to the digit displayed on the screen.

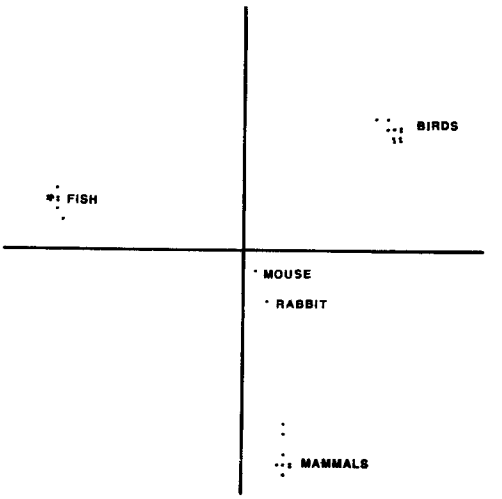

FIG. 1. Multidimensional scaling solution from similarity ratings (dimensions 1 and 2).

Results and Discussion

Individual differences in ratings among subjects were analyzed by an inverted principal-components factor analysis before obtaining a composite dissimilarity matrix. The first component, indicative of communality among subjects, accounted for 70% of the total variance, leaving only 30% for stable individual differences and subject unreliability. In addition, the subjects did not appear to fall into any obvious subgroupings. It was therefore deemed appropriate to average over the 15 subjects to obtain a composite matrix.

This composite distance (dissimilarity) matrix was then analyzed by a nonmetric multidimensional scaling model using TORSCA (Young & Torgerson, 1967). Euclidian solutions of varying dimensionality were tried, and a four-dimensional configuration appeared adequate. The stress (Kruskal, 1964) was .031, while the index of agreement (Torgerson & Meuser, 1962) was .9997, indicating a close monotonic fit of the four-dimensional configuration to the data. Even within categories, the product-moment correlation between the 135 ratings of dissimilarity and the derived distances in the space was .93.

The results were rotated to maximize the contribution of class membership to the structure of projections in a two-dimensional subspace. Figure 1 shows the configuration of the projections of the stimuli in this subspace. Clustering due to the class membership is clear. It is interesting to note, however, that rabbit and mouse seem to form a minicluster somewhat distinct from the other mammals used in the study. Although it is possible that these species actually form a separate semantic class, it is also likely that the gap between the two mammal clusters simply resulted from the limited selection of mammals used in the study. Had mammals such as fox, muskrat, and skunk been included, the gap might well have disappeared. A definitive resolution would require further experimentation. The analyses used in this investigation treat rabbit and mouse as members of the same class as the other mammals.

Figure 2 shows dimensions 3 and 4, the quantitative variation. As with the previous scaling of mammals (Henley, 1969; Rips et al., 1973) the two salient features appear to be *size* and *ferocity*. It is interesting to note that the bird and fish terms do not appear to vary nearly as much as do the mammal terms. Table 2 shows the means and standard

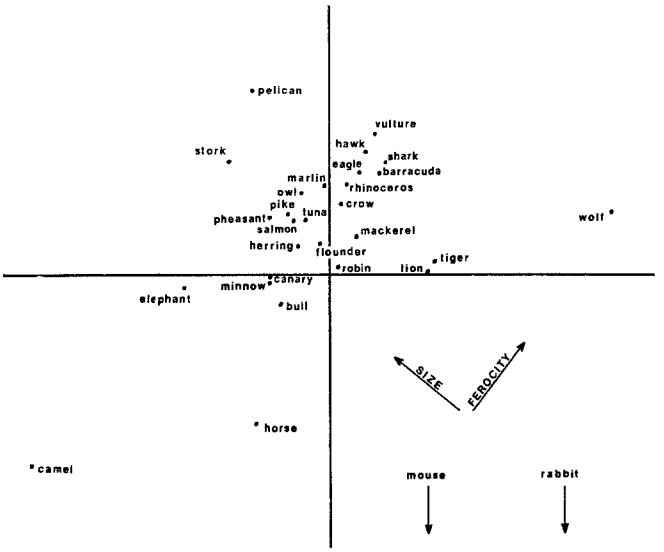

FIG. 2. Multidimensional scaling solution from similarity ratings (dimensions 3 and 4).

TABLE 2

DISSIMILARITY RATINGS WITHIN AND BETWEEN CATEGORIES

Group	Mean	*SD*
Mammals	4.62	1.51
Birds	2.58	.75
Fish	2.00	.67
Between	6.97	.71

deviations for interelement ratings within and between groups. The mean dissimilarity between mammals was considerably greater than between birds or fish. The standard deviations showed a considerable increase in variability for ratings within the mammal category. Thus, although the words were matched for frequency of occurrence across categories, the mammals included appear to be much more dissimilar than birds or fish. Whether this difference is relative only to the task at hand or is absolute cannot be answered at this time.

A plot of the judgment latencies as a function of rated similarity indicated that the slowest judgments corresponded to the middle of the rating scale. Since this effect might have been due to the search process, each subject's mean times for the reaction time task were subtracted from their judgment latencies according to the appropriate key position. A logarithmic transformation was then applied to the data. The transformed data was then averaged across subjects. Figure 3 is a graph of the antilogs of the mean corrected log latencies as a function of rated similarity.

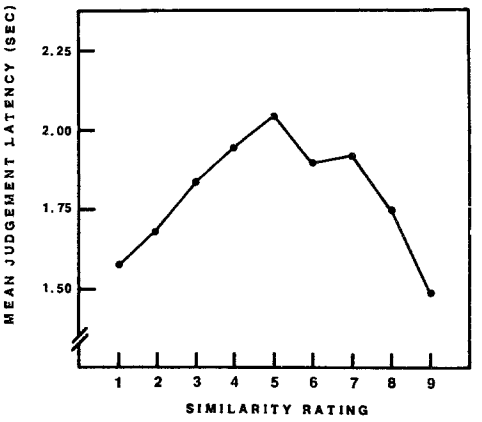

FIG. 3. Similarity judgment latency as a function of rated similarity.

Correction for key position only minimally reduced the curvature of the graph. It thus appears that pairs rated either very similar or very dissimilar were relatively easy to rate, while pairs which were marginally similar were the most difficult to characterize.

The log latencies were then dichotomized according to whether the judgments were made within categories or between categories. Essentially all ratings within categories were between 1 and 5; most all ratings between categories were between 5 and 9. The product-moment correlation between log latency and rated distance within groups was .653; across groups the correlation was comparable but of opposite sign, −.612. The present experiment involves two different kinds of variation: a nominal one corresponding to class membership and a quantitative kind allowing continuous variation. Stimuli that are similar with respect to both types or very different with respect to both kinds of variation have short response times. Longer response times are associated with stimulus pairs for which the two types of variation are in conflict—either same class but very different on the quantitative dimensions or different classes but similar on the quantitative dimensions.

One might argue that the within category effect is due to an activation in memory retrieval (Meyer & Schvaneveldt, 1971); however, the latency differences are greater (by a factor of 10) than could reasonably be attributed to differential retrieval times. Moreover, the activation effect would not account for the inverse relation found for cross-category judgments. Several investigators (e.g., Schaeffer & Wallace, 1970) have found that subjects take longer to reject a false proposition if the subject and predicate are related, than if they are unrelated. However, in such a paradigm, the decision to accept or reject the proposition was strictly a logical matter. In the study at hand the logical structure of the domain was not emphasized, rather subjects were told to rate on the basis of their immediate and overall feeling about the similarity between the two words of each pair. Given such a paradigm, one might not *a priori* suspect that judgment latencies would be a function of the directly judged similarity.

Experiment I demonstrates that the relations both within and between related categories can be represented in a multidimensional space, where the interelement distances are interpreted as semantic relatedness. It also shows that the ease of comparing two words can be represented as a function of their semantic relatedness.

Experiment II: Semantic Structure and Categorization

Rips et al. (1973) found that derived distances between points in the subjective structures did not predict categorization time as well as did raw similarity ratings. Although they did not give the values of stress obtained from their scalings, the correlations between ratings and derived distance (.68 for *birds*, .65 for *mammals*) indicate that their derived space simply did not reflect the relations present in the raw data in a precise way.

Another problem with the Rips et al. study was that typicality and similarity were completely confounded. Typicality refers to the grade of membership of a subordinate in a superordinate category (Lakoff, 1972) while similarity ordinarily refers to the communality of two elements at the same level in the hierarchy. By attempting to fit similarity ratings (e.g., *lion–mouse*) and typicality ratings (e.g., *rabbit–mammal*) into the same spatial representation, both relations became distorted. Thus, while the space had a certain intuitive reasonableness, it simply did not fit the data.

Given the close fit between the data and the space in Experiment I (stress = .031) and the fact that the space accounted for variability both within and between categories, it was desirable to determine in what way (if any) the subjective structure related to RT in a same-different categorization task.

Method

Subjects. Sixteen students at The Johns Hopkins University, who had not been tested in Experiment I, served as paid subjects.

Procedure. The computerized procedure used in Experiment I was repeated here. Subjects were instructed that 30 animal terms, 10 from each of the three categories, *birds*, *fish*, and *mammals*, would be presented on the screen, two at a time. If both words were members of the same category, a "same" response was indicated. If the two words were from different categories, a "different" respone was appropriate. Response keys were counterbalanced across preferred hands. A single trial consisted of the following. A stimulus pair would appear on the screen and remain on until the subject responded. One second after the subject's response the next stimulus pair appeared on the screen. A rest period of five minutes was given halfway through the session. The entire experimental session lasted approximately 50 minutes. Subjects were instructed that they should respond as fast as was consistent with insuring correct responses. Ten practice trials with words from the three categories but not on the list of 30 preceded the experimental session.

Results and Discussion

Errors were summed across subjects to obtain an error matrix relating the number of incorrect responses to the appropriate stimulus pairs. It was felt that since the total error rate was low (4.2%), and since over 50% of the cells in the error matrix were empty, no reliable conclusions could be drawn from an error analysis. However, one interesting aspect of the incorrect responses was that subjects incorrectly identified animals as coming from different categories over 40% less often for mammals than for fish or birds (31 errors vs. 54 and 56 respectively). The more accurate performance with mammals and the larger degree of dissimilarity might have been the result of the relatively greater familiarity subjects showed with the concepts in this category.

In Experiment I it was found that judgment latencies were related to the degree of similarity of members of the word pairs. If *RT*s were likewise influenced by similarity then one would expect a correlation between the relatedness of the members of each stimulus pair and the time necessary to categorize the two words. Log reaction times for correct responses were summed across subjects to arrive at a composite reaction time matrix. The data were then dichotomized according to whether the animal pair was from within the same category or different categories.

Due to the goodness of fit of the derived distances to the similarity ratings, there were no significant differences in correlations between those obtained with derived distances or the raw ratings (cf. Rips et al., 1973, p. 10). Thus while only the correlations involving the similarity ratings will be discussed, those obtained using the derived distances were essentially the same.

When log *RT* was correlated with ratings across the 135 pairs within the three categories, no relation was found, $r = .15$, $p > .05$. The lack of a relation was a result of the relative differences in ratings between mammals and the other categories. When the ratings for the three groups were equated in mean and standard deviation a moderate relation was found between rating and log *RT*, $r = .45$. The corresponding correlation for the 300 pairs across categories was negative and marginally significant, $r = -.17$, $p < .05$.

When the within-groups judgments were correlated with log *RT* separately for each category, the resulting correlations were .49, .24, and .61 for mammals, birds, and fish, respectively. Although these correlations are in the same range as found by Rips et al. the important factor is that a moderate relation appears to exist between the *RT*s and the semantic distances as derived from a procedure where terms from three separate categories were rated simultaneously. These results

imply that subjects do operate with comparable subjective structures across different categories.

The moderate correlations do indicate that there is some relation between the construct of semantic distance and the time required to decide whether two animals belong to the same category.

Experiment III: Typicality Judgments

Since the metric of semantic distance or similarity rating accounted for such a small proportion of the variability in the reaction time task, perhaps the typicality of an instance to its category concept might show a more direct relation. Typicality has been proposed as a parameter which influences the time needed to decide on category membership (Smith et al., 1974). Accordingly, in the next experiment typicality ratings were derived for the words under consideration. The ratings were then viewed in reference to the results of the previous experiment.

Method

Subjects. Thirty-eight undergraduates at Towson (Md.) State College served as volunteer subjects for this study.

Procedure. Subjects were given a sheet containing three columns representing the categories used in the study. In each column the ten exemplars were listed, followed by a blank space. Subjects were instructed to first look at the column of mammals, and to enter a "1" in the space beside the word which was most typical of the category name. Subjects were then asked to locate the word in the list that corresponded to the least typical instance of the category and to give that word a rating of "10". The remaining words in the list were to be assigned numbers corresponding to their typicality using the first two words selected as anchor points. The procedure was repeated for the fish and birds.

Results and Discussion

The typicality ratings were summed over subjects. The resulting mean ratings are shown in Table 3 along with the frequency of occurrence from the Battig and Montague norms. As has been found in other studies, the typicality ratings correlate quite highly with Battig and Montague conjoint frequencies (−.79, −.85, and −.47 for mammals, birds, and fish). The negative correlations resulted from high typicality corresponding to low numbers.

In the Rips et al. (1973) study, the category name in the multidimensional solution appeared as approximately the centroid of the space. The derived distance from the position

TABLE 3

Mean Typicality Ratings and Production Frequency

Mammals			Fish			Birds		
Term	Typicality	Frequency	Term	Typicality	Frequency	Term	Typicality	Frequency
Horse	2.95	348	Tuna	2.65	139	Robin	1.65	377
Lion	3.03	225	Salmon	2.81	142	Canary	2.35	134
Elephant	3.27	182	Flounder	3.62	50	Eagle	3.22	161
Tiger	3.46	203	Mackerel	4.35	27	Crow	3.84	149
Wolf	4.68	55	Herring	4.41	161	Owl	4.89	36
Rabbit	5.30	48	Marlin	4.51	33	Hawk	5.22	111
Camel	5.78	28	Shark	4.62	176	Vulture	6.59	44
Bull	6.16	40	Minnow	5.41	54	Pheasant	6.86	22
Rhinoceros	6.97	31	Barracuda	6.24	31	Stork	7.51	10
Mouse	7.08	118	Pike	6.54	58	Pelican	8.24	11

of the exemplar to the category name was then used to predict *RT*. In the experiment under consideration the typicality ratings could also be considered as distances from the centroid represented by the category name (Rosch, 1973). The actual centroid for each category in the multidimensional space might also be considered as representing the prototype of the category or the category concept exemplified by the name. Both concepts, that of the typicality ratings considered as distances to the centroid (rated centroid) and that of the calculated distance to the centroid of the derived space (derived centroid), were used to calculate multiple correlations between the log reaction times from experiment two and the distances from each exemplar to its respective centroid. The multiple correlations were computed separately for each category. The results are presented in Table 4, along with the correlations from the similarity ratings. It was clear that neither method of determining typicality led to substantial improvements in prediction.

TABLE 4

PRODUCT-MOMENT CORRELATIONS BETWEEN LOG *RT* AND SEVERAL SEMANTIC DISTANCE MEASURES

	Distance measure		
Group	Similarity rating	Rated centroid	Derived centroid
Mammals	.49	.61	.51
Birds	.24	.46	.42
Fish	.61	.65	.66

EXPERIMENT IV: SEMANTIC STRUCTURE AND FREE RECALL

Since there appears to be some relation between semantic distance and performance on the categorization task, it was felt that the structure represented by the multidimensional solution should also influence performance in a memory task. Performance on free recall tasks obviously depends on a large number of variables. Of these the semantic structure of the words to be recalled would be expected to be of substantial importance. A methodologically important distinction, however, exists between the tasks used in the first experiment and the present experiment. Thus the judgment task in Experiment I allowed each subject to reflect in his response the potential multidimensional structure underlying the original space. In a recall task, such as the present one, on the other hand, the nature of the task constrains the types of structures to strictly linear ones. That is, the only retrievable information present in the response protocol of a subject is a unidimensional ordering of the original terms. This limitation can be surpassed if the assumption is made that in averaging across subjects we can approximate an "ideal" structure common to all speakers of that language. Thus variations among subjects do not reflect different underlying linear orderings but rather different emphasis of selected dimensions of a multidimensional structure. This is a commonly made assumption in clustering experiments (Miller, 1969).

The next experiment was carried out to examine the influence of semantic structure on free recall for the same set of stimuli as used in the previous experiments.

Method

Subjects. Twenty-four students at The Johns Hopkins University served as volunteer subjects for this study.

Materials. Six randomized lists of the thirty stimulus words were recorded on magnetic tape with an interstimulus interval of 2 sec. Instructions to the subjects were also recorded at the beginning of the tape.

Procedure. The experiment was run with groups of subjects ranging from one to six. Subjects were informed that they would hear a 30-word list at a rate of one word every 2 sec. They were further instructed to listen carefully and to recall as many words as they were capable of at the end of the list. Subjects were given 2 min to write down as many words as

they could remember on a response sheet supplied to them. The response sheet was then removed and the process was repeated for five more random lists.

Results and Discussion

As pointed out earlier using the free recall procedure we obtain only a linear sequential ordering for the animals on a given trial for each subject. Hence any more complex structure cannot be revealed on a single trial for a single subject. A modification of Friendly's proximity analysis (Friendly, 1972) was used to resolve the problem. Friendly's original procedure arrives at a proximity (or conversely, distance) matrix for the terms in the recall lists based on the assumption that words that are highly related in memory will consistently tend to be recalled contiguously; and thus the interitem distance in the recall list will be an indicant of the degree of relatedness between the two items. In this procedure a dissimilarity matrix is obtained by averaging *over trials.*

In contrast, in the present experiment following the assumption stated in the introduction, we averaged *over subjects* for single trials instead. The results of Experiment I indicated that the subjects could be considered to have essentially the same semantic structure. It therefore seemed appropriate to obtain a dissimilarity matrix for each of the six recall trials by averaging over subjects.

To obtain an indication of the relation between the similarity ratings and the interitem distances from the recall lists, correlations were calculated for each category on each of the six recall trials. The results from the first and sixth trials are shown in Table 5. Overall, there was some degree of correspondence between the similarity ratings and the closeness of the words in the recall lists, and the relation appeared to improve from trial to trial.

To further analyze the structure in the recall data, a multidimensional scaling was performed on the distance matrices from trials one and six. To be consistent with the solution from the similarity data, four-dimensional solutions were obtained using TORSCA. Stress values were .159 and .088 for the first and sixth trials, respectively. Figures 4 and 5 show the rotated solutions for the first and second dimensions in each scaling. Very clearly these dimensions represent the clustering due to the three categories. This discrete group structure is most obvious in the solution for the last trial; however, it is already present after only one exposure to the stimuli.

TABLE 5

PRODUCT-MOMENT CORRELATIONS BETWEEN SIMILARITY RATING AND FREE RECALL PROXIMITY ($N = 45$)

	Recall trial	
Group	Trial 1	Trial 6
Mammals	.362	.554
Birds	−.020[a]	.395
Fish	.113[a]	.163[a]

[a] Not significant at .05 level.

While the first and second dimensions of the TORSCA solutions compared favorably with the solution from the similarity ratings, the third and fourth dimensions were uninterpretable. Perhaps a spatial representation was not appropriate. To examine the relations within the data from another vantage point, a clustering solution was obtained using HCS, Hierarchical Clustering Scheme (Johnson, 1967).

Figure 6 shows the hierarchical structure from trial six of the recall task. For comparison a similar analysis was performed on the similarity ratings. The rating structure is represented in Fig. 7. There is considerable similarity between the two structures. Closely clustered groupings such as *lion–tiger*, *eagle–hawk–vulture*, *salmon–tuna*, and *elephant–rhinoceros* appear in both hierarchies. The three discrete categories also are discriminated in both cases.

It should be noted, however, that the correspondence is not exact. *Pelican* and

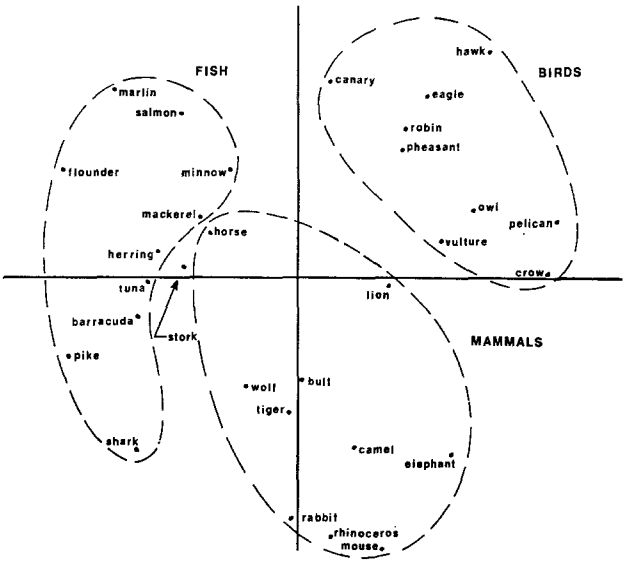

FIG. 4. Multidimensional scaling solution from free recall proximity matrix of trial one (dimensions 1 and 2).

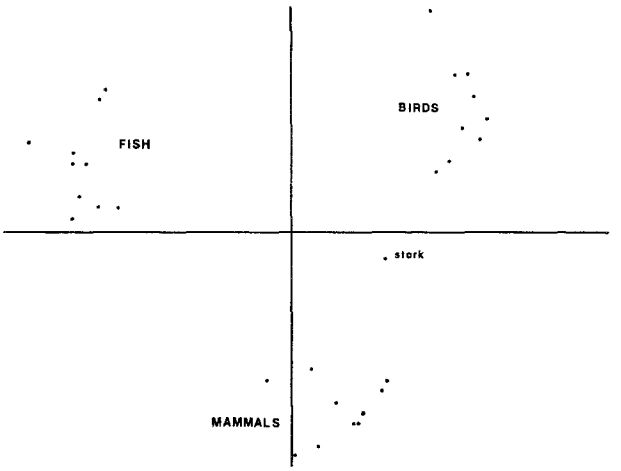

FIG. 5. Multidimensional scaling solution from free recall proximity matrix of trial six (dimensions 1 and 2).

pheasant were closely clustered in the recall structure but not in the similarity hierarchy. Also in the recall configuration, the three food fish of *tuna*, *salmon*, and *mackerel* combined with the mammal group before connecting to the remaining seven fish. Other similarities and differences can be detected throughout the two solutions. While the two procedures of similarity rating and free recall result in similar category (nondimensional) configurations, the differences between the structures imply that differential retrieval and organizational processes might be operating in the two tasks.

The individual recall protocols were helpful in analyzing the relations within the retrieval process. Unusual transitions (i.e., in terms of the derived similarity structure) such as *elephant–mouse* and *pelican–herring* were not uncommon. Intrusions such as *falcon* (which was not one of the stimulus words) following *eagle* were also observed. It appears that all

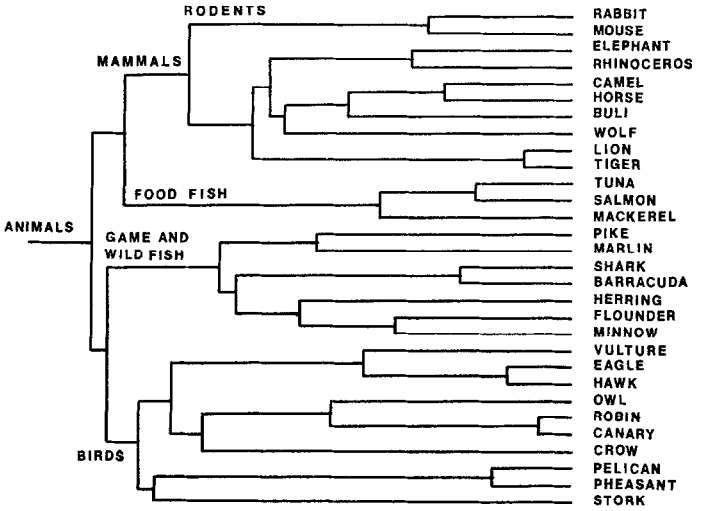

FIG. 6. Hierarchical clustering solution from proximity matrix of sixth recall trial.

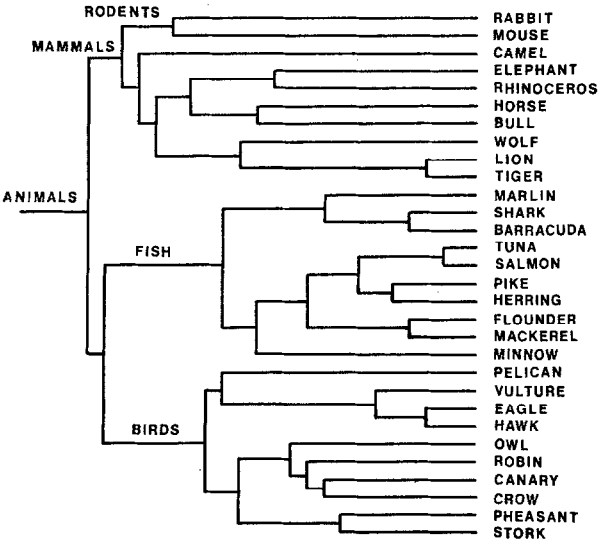

FIG. 7. Hierarchical clustering solution from similarity ratings of Experiment I.

available information was employed in recalling the terms. Subjects have had a lifetime of exposure to the words used in the study. If asked, they could certainly produce considerably more information on all the animals than simply size, ferocity, and group membership. The use of this information was constrained in the categorization task. In the free recall experiment, any available information could have been used.

GENERAL DISCUSSION

It is obvious from the results obtained in our experiments and those cited in the introduction that semantic distance can be a powerful predictor of performance in tasks requiring the use of information from semantic memory. It should be equally obvious, however, that in many tasks the amount of variance attributable to variations in semantic

distance can be extremely small as was the case in the *different* judgments of Experiment II where the correlation between log *RT* and semantic distance was only −.17 (or 3% of the total variance). Thus, we must conclude that the predictor value of semantic distance may vary both as a function of tasks and categories tested.

In the similarity ratings it was found that mammals were subjectively more dissimilar than either the fish or the birds. In addition the mammal category produced fewer errors in the categorization task. While the words in each group were matched for frequency of occurrence, subjects' experiences with the words still varied over a large range. In the subject population under consideration (college students) much more is typically known about the mammals than either the birds or fish we tested. Not only are elephants large and placid, but they are gray, found in circuses, love peanuts, and are scared of mice. Little more is known about herring than that they are relatively small fish which live in the ocean and are great pickled with sour cream. In general not much more is known about the fish or bird categories than the dichotomy between the groups and a little about the variables of *size* and *ferocity*. Thus, it may be argued, for example, that the large discrepancy observed among the correlations between semantic distance and free recall proximity for the three animal categories is due to the relative differences in familiarity with these categories. This line of reasoning leads to the position that (as far as the free recall task is concerned) the dimensions of *size* and *ferocity* are psychologically salient, that is, can be used in organizing a response, only for the category *mammal* (and perhaps weakly for *birds*).

Another important discrepancy is the difference in the size of correlations obtained for the first and second experiment. Recall that in Experiment I the correlations between judgment latencies and rated similarity for within and between categories were .65 and −.61, respectively. In Experiment II the comparable correlations were .45 and −.17. Thus, for within-category judgments there is a moderate relationship between semantic distance and response time for both Experiments I and II. For between-category judgments, however, a moderate correlation was obtained only in Experiment I. In Experiment II a statistically significant but otherwise unimportant relationship obtained between semantic distance and *RT*. From this pattern of results we may conclude the following.

First, the semantic distance between pairs of items within a category and between categories contributes significantly to decisional complexity, as measured by response time, in determining degree of similarity. Thus, the further inference is justified that the dimensions of *size* and *ferocity*, in our specific case, are psychologically saleint in rating animal pairs on similarity. Second, a similar conclusion to the above can be reached for the *same* judgments (within category pairs) in the categorization task. That is, semantic distances within category, which we interpreted as reflecting the size and ferocity of the animals, account for a substantial part of the variation in judging two items are members of the same category. For the *different* judgments (between category pairs), however, semantic distance was not an important determinant of *RT*. This can be interpreted as suggesting that in *different* judgments variations along the quantitative dimensions of *size* and *ferocity* are relatively unimportant compared to category membership.

A further implication of our results is that different kinds of semantic information are used as a function of experimental task. Thus in the first three experiments the important variable was the relation holding among objects in a semantic field. The determining factor influencing a subject's response is the degree to which two words share components of meaning relevant to class membership—structural meaning. In the fourth experiment, an important part could be played by associative meaning: relations holding among words

which, strictly speaking, are not relevant to a definition of the words. Thus a subject may recall two words contiguously due to some accidental connection between those two words. Both structural and associative meaning mediate free recall.

In our discussion we have omitted any mention of how our results relate to current views and models of semantic memory. The reason is that these studies were not strictly designed to test any explicit model of semantic memory. Rather our aim was simply to explicate the relevance of the construct semantic distance in various tasks for a richer and more structured subset of the lexicon than previously investigated. And, in this context, our data yielded a number of general statements. Overall, subjectively derived semantic distance is a good predictor of performance in timed tasks and free recall tasks when retrieving information from semantic memory. This general conclusion should be tempered, however, by the fact that the usefulness of semantic distance as a predictor depends on: the type of task used and on the relative familiarity of the material tested.

References

ARNOLD, J. B. A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology Monograph*, 1971, **90**, 2, 349–372.

BATTIG, W. F., & MONTAGUE, W. C. 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, Part 2).

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.

FILLENBAUM, S., & RAPOPORT, A. *Structures in the Subjective Lexicon*. New York: Academic Press, 1971.

FRIENDLY, M. L. *Proximity analysis and the structure of organization in free recall*, Techn. Report RB-72-3, Princeton: ETS, 1972.

GLASS, A. L., & HOLYOAK, K. The effect of Some and All on reaction time for semantic decisions. *Memory & Cognition*, 1974, **2**, 436–440.

GLASS, A. L., HOLYOAK, K., & O'DELL, C. Production frequency and the verification of quantified statements. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 237–254.

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

JOHNSON, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, **32**, 241–254.

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

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.

LUCE, R. D. *Individual choice behavior*. New York: Wiley, 1959.

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

MEYER, D. E., & Schvaneveldt, R. W. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 1971, **90**, 227–234.

MILLER, G. A. Psycholinguistic approaches to the study of communication. In D. L. Arm (Ed.) *Journeys in science: small steps-great strides*. Albuquerque: University of New Mexico Press, 1967.

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 semantics and lexical memory. In A. W. Melton and E. Martin (Eds.) *Coding processes in human memory*. Washington, D.C.: V. H. Winston and Sons, 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.

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

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.

SMITH, E. E., SHOBEN, E. J., & RIPS, L. J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 1974, **81**, 214–241.

TORGERSON, W. S. Multidimensional scaling of similarity. *Psychometrika*, 1965, **30**, 379–393.

TORGERSON, W. S., & MEUSER, G. Informal notes on Torgerson and Meuser's IBM 7094 Program for Multidimensional Scaling. Cambridge: Mitre Corp., 1962. Mimeographed Report.

YOUNG, F. W., & TORGERSON, W. S. TORSCA, A FORTRAN IV Program for Shephard-Kruskal multidimensional scaling analysis. *Behavioral Science*, 1967, **12**, 498.

ZURIF, E. B., CARAMAZZA, A., MYERSON, R., & GALVIN, J. Semantic feature representations for normal and aphasic language. *Brain and Language*, 1974, **1**, 167–187.

(Received January 13, 1975)