Bulletin of the Psychonomic Society
1984, 22 (5), 377-380

Reliability of retrieval from semantic memory: Noun meanings

FRANCIS S. BELLEZZA
Ohio University, Athens, Ohio

The purpose of this experiment was to determine how reliably college students could retrieve from memory semantic information about commonly used nouns. In two recall sessions separated by 1 week, subjects gave definitions for abstract nouns, category labels, and concrete nouns. It was found that more propositions were given in definitions of concrete nouns (7.13) than in definitions of category labels (6.73) or abstract nouns (4.13). The within-subject reliability of retrieval was .55 for concrete nouns, .46 for category labels, and .43 for abstract nouns. The between-subjects reliability was .29 for concrete nouns, .20 for category labels, and .17 for abstract nouns. Discussed are the implications of these data for the hypothesis that the single word represents the unit of meaning.

The agreement among people about the meaning of words indicates *inter*subjective reliability and ensures that effective communication can take place. Discussion of the breakdown of this sharing of meanings is commonplace, and much has been written about what different people mean by "independence," "responsibility," "democracy," "opportunity," and so on (Ogden & Richards, 1923; Wittgenstein, 1968). However, *intra*subjective reliability of meaning is also important and necessary for intersubjective reliability to occur. The network of relations among concepts lending meaning to a word in memory must be relatively stable from day to day, and this network must be used in a similar manner from day to day. Intrasubjective reliability of meaning is a prerequisite for intersubjective reliability.

The distinction betwen episodic and semantic memory implies that information in semantic memory is relatively stable (Tulving, 1972). This particular assumption, however, has not been widely tested. Because semantic memory can be considered as an information storage and retrieval system (Lindsay & Norman, 1977, chap. 10), questions can be raised about how reliably information in semantic memory can be retrieved. The results of Bellezza (1984) indicate that instances of common categories are not retrieved from semantic memory in a completely reliable manner. A degree of unreliability of retrieval may be true of category instances, but is it true for the retrieval of verbal attributes of words such as familiar nouns? To try to answer this question, subjects were asked to give definitions for concrete nouns, category labels, and abstract nouns. A week later, they performed the same task for the same words, and the content of the two sets of definitions obtained from each subject was compared.

Portions of this research were presented at the meeting of the Psychonomic Society in Minneapolis, Minnesota, in November 1982. This research was supported in part by a grant from the Field-Wiltsie Foundation. Thanks go to Kathy Kamin, Sylvia Sims, Rosalind Whatley, and Cathy Young for their assistance in collecting and scoring the data. Thanks also go to Ohio University Computer and Learning Services for making computer time and their facilities available. Requests for reprints should be sent to Francis S. Bellezza, Department of Psychology, Ohio University, Athens, OH 45701.

METHOD

Subjects

Thirty-three undergraduates from introductory psychology courses at Ohio University volunteered to participate for extra course credit.

Materials

Eight abstract nouns and eight concrete nouns were selected from the Paivio, Yuille, and Madigan (1968) noun list. The concrete nouns had concreteness ratings that varied between 6.55 and 7.00, and the abstract nouns had concreteness ratings that varied between 1.46 and 3.62. The median general frequency of occurrence for both sets of nouns was between 50 and 100 per million words (Thorndike & Lorge, 1944). The eight concrete nouns used were apple, car, cat, chair, doll, gun, hammer, and robin. The eight abstract nouns used were advantage, chance, concept, essence, instance, origin, outcome, and situation. Also, eight category labels were chosen from Battig and Montague (1969). These were clothing, dwelling, fish, flower, insect, stone, tree, and vegetable.

Procedure

In the first session, the subjects were given a booklet containing the 24 words listed in a random order. The 20 subjects in the sentence condition were asked to define, in full sentences, each word in the list as well as they could. The example for the word "dog" was "A dog is a small four-footed mammal often found as a household pet. It likes to eat meat, bark, and wag its tail." The example given for the word "idea" was "An idea is a mental event that a person experiences and that sometimes indicates what that person should do next. It is important to have good ideas." Pilot work indicated that scoring these definitions for propositions (Kintsch, 1972) was possible, but difficult. Therefore, a second group of subjects were tested in the phrase condition and were instructed to give definitions of one or two words rather than complete sentences. One- and two-

word phrases can represent attributes and predicates of words (Lindsay & Norman, 1977, chap. 10; Kintsch, 1972; Moore & Newell, 1973). The example given for "dog" was "animal, pet, has fur, growls, wags tail, bites, eats meat, barks, is small." The example given for the word "idea" was "mental event, person experiences, solves problems, important." All subjects were allowed 1 min to define each noun and were paced through the list by the experimenter. One week later, the subjects returned and were given the same instructions and words to define, except that the words were in a different random order.

RESULTS

The definitions provided by subjects in the phrase condition were scored in the form that they were given by the subjects. However, the definitions provided by the subjects in the sentence condition were each analyzed into a series of propositions (Kintsch, 1972). For example, for the word "hammer," one definition given was "A tool used to pound nails or to take out nails. It usually has a wooden handle and a metal or steel top." The eight propositions derived from this definition were: A hammer is a tool. A hammer is used to pound nails. A hammer is used to take out nails. A hammer has a handle. The handle is wooden. A hammer has a top. The top is steel. The top is metal. For the word "fact" a definition given was "Not false. Something that is true and actually happened or will happen." The propositions derived were: A fact is not false. A fact is true. A fact actually happened. A fact will happen.

Number of Attributes

For the subjects in the phrase condition, the number of one- and two-word phrases in each definition were counted. For the subjects in the sentence condition, the simple propositions each definition comprised were counted. The mean number of propositions or phrases given by the subjects in each condition for each of the three types of nouns is shown in Table 1. A 2 x 3 x 2 analysis of variance was performed on the mean numbers of phrases or propositions recalled. The factors were response condition (phrase vs. sentence), noun type (abstract, category, concrete), and session (Session 1 vs. Session 2). Only response condition was a between-subjects factor. Noun type was significant [$F(2,62) = 177.36$, $p < .001$, $MSe = .994$]. With a priori contrasts, it was found that more propositions were given for concrete nouns (7.13) than for category nouns (6.73) [$F(1,62) = 5.23$, $p < .05$]. Also, more propositions were given for category nouns (6.73) than for abstract nouns (4.13) [$F(1,62) = 224.27$, $p < .001$]. There were some additional significant effects. More propositions were given in Session 1 (6.17) than in Session 2 (5.82) [$F(1,31) = 5.58$, $p < .025$, $MSe = .755$]. There was also a noun type x session interaction [$F(2,62) = 3.48$, $p < .05$, $MSe = .494$]. A Cicchetti (1972) test showed that there was a significant difference between the number of propositions recalled for concrete nouns in Session 1 and Session 2, but there was no significant difference for category and abstract nouns. This can be seen in Table 1. Finally, there was no significant effect of response condition, but there was a significant response condition x noun type interaction [$F(2,62) = 3.26$, $p < .05$, $MSe = .995$]. A Cicchetti test on the means, however, detected no differences. This failure to detect differences using posttests probably resulted from the fact that the interaction was only marginally significant.

Within-Subject Reliability

On the basis of the common-element correlation (Bellezza, 1984), mean overlap scores for each noun type were computed for each response condition. These means are shown in Table 1. The mean overlap score for each noun type was computed for each subject, and a 2 x 3 analysis of variance was performed with the factors response condition and noun type, respectively. Only response condition was a between-subjects condition. The only significant source of variation was noun type [$F(2,62) = 12.18$, $p < .001$, $MSe = .009$]. Tukey's HSD tests (Kirk, 1968) showed that the difference between the means for concrete nouns and category nouns was significant, but that the difference between the means for category nouns and abstract nouns was not.

Between-Subjects Variability

To determine the degree of agreement among subjects concerning the definitions, the subjects were randomly paired and a common-element correlation was computed

Table 1
Mean Recall and Correlation Values for Concrete, Category, and Abstract Nouns Used in the Sentence and Phrase Conditions

Condition	Number of Propositions Recalled in Session 1	Number of Propositions Recalled in Session 2	Within-Subjects Overlap	Between-Subjects Overlap
Sentence				
Concrete	7.79	7.46	.56	.31
Category	7.12	6.95	.46	.19
Abstract	3.94	4.23	.44	.17
Phrase				
Concrete	7.24	6.37	.53	.26
Category	6.80	6.26	.46	.21
Abstract	4.24	4.08	.42	.19

for each word in each pair. For the 20 subjects in the sentence condition, 10 pairs were formed from the data of Session 1 and 10 different pairs were formed from the data of Session 2. The same procedure was used for the 13 subjects in the phrase condition, except that only 6 pairs could be formed for each session. A 2 x 2 x 3 analysis of variance was performed on the mean overlap scores formed from the abstract, category, and concrete nouns. The factors were definition condition (sentence vs. phrase), session (Session 1 vs. Session 2), and word type (abstract, category, concrete). Only noun type was a within-pairs factor. Only two sources of variation were significant. The overlap scores from Session 2 were significantly greater than the overlap scores from Session 1 [$F(1,28) = 5.97$, $p < .025$, $MSe = .006$]. The mean overlap scores were .20 for Session 1 and .24 for Session 2. There is no obvious explanation for this difference. The other significant source of variation was noun type [$F(2,56) = 24.04$, $p < .001$, $MSe = .004$]. The mean overlap scores were .17 for abstract nouns, .20 for category nouns, and .29 for concrete nouns. Posttests using Tukey's HSD procedure showed that the overlap score for the concrete nouns was significantly larger than that for the abstract nouns and category nouns. These latter two means were not significantly different from one another.

DISCUSSION

The major result of this experiment was that the mean overlap score for all three types of nouns combined was .48, with a range in value of .43 for the abstract nouns to .55 for the concrete nouns. These results represent a rather modest reliability of retrieval of noun meanings from semantic memory. It is possible to argue that, because a single word is not a more effective prompt for the retrieval of information from semantic memory, then that single word has no well-defined meaning for the person using it. If a word does have a well-defined meaning, then that meaning should be the same from one occasion of use to the next. But this does not occur. The modest reliabilities found when common nouns are defined at different times suggests that the single word should not be considered a valid unit of meaning. On the basis of this criterion, a more valid unit of meaning may be the proposition (Anderson, 1981). Whether propositions used as prompts result in more reliable retrieval from semantic memory is an empirical question, however, and the reliability of information retrieved from semantic memory relevant to propositions has yet to be measured.

One could argue that many words, including the nouns used here, are polysemous (Clark & Clark, 1977, pp. 444-446) and that different meanings were given each week to the presented words. This could reflect the fact that the words indeed do have different meanings. Inspection of the data indicates that very few of the overlap scores were 0 or 1. In fact, only .08 of the overlap scores had values of 0, and only .03 had values of 1.

Another objection might be that the definitions were the same, but that the subjects expressed them in different words. For example, for the word "dog," the first definition may have included "barks," but the second definition included "makes noise." Another possibility is "sheds fur" versus "loses hair." Again, inspection of the definitions showed that very few were instances of the same attributes or propositions expressed in different words in the two sessions. The difference in the two definitions given for any word represented differences in the ideas expressed and did not represent simply the same ideas expressed in different words. To check for any differences in a systematic fashion, the overlap scores of the 13 subjects in the phrase condition were rescored to find phrases in the two sessions for each noun-subject combination that were similar in meaning but different in wording. Few were found, and the recomputed overlap scores for the abstract, concrete, and category nouns were unchanged.

The results are also important for methodological reasons. Definitions comprising one- and two-word phrases gave results that were almost identical to the results for full sentences analyzed into propositional components. This is important because the sentence data required a great deal of time to score, whereas the one- and two-word phrases were scored more easily.

As expected from the results of Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), the mean number of propositions recalled for the concrete nouns was greater than the number recalled for the category nouns. This should occur because each concrete noun represents a fairly specific class of physical objects, and, therefore, propositions representing sensory-perceptual information can be recalled for these nouns. There exists less specific sensory-perceptual information for category nouns and even less for abstract nouns. Therefore, the number of propositions retrieved for abstract nouns was the smallest.

Also, a larger overlap score was obtained for the concrete nouns than for the category nouns or the abstract nouns. This result is somewhat surprising when compared with the results of Bellezza (1984), which indicated that the number of instances retrieved for a category was inversely related to the value of the overlap score. Bellezza suggested that for a large category a particular item is less likely to be retrieved during both sessions. However, in the present experiment, both a greater number of propositions and greater overlap scores were obtained for concrete nouns than for the other two types. To explain this, it could be argued that the larger overlap scores for the concrete nouns reflect the finding that high-imagery nouns are better recall cues than are low-imagery nouns (Paivio, 1971). In Session 2, the subjects may not only have tried to define each noun, but may also have tried to remember their definitions from the previous week. This hypothesis could also account for the result that fewer propositions were generated during Session 2 than during Session 1, especially in the case of concrete nouns. During Session 2, the subjects may have spent time trying to remember what definitions they produced during Session 1, even though they were instructed only to define the nouns.

Between-Subjects Reliability

The mean between-subjects overlap score was lower for the abstract nouns and category nouns than for the concrete nouns. Not only did the subjects define concrete nouns from week to week more similarly than they did abstract nouns, but different subjects also defined concrete nouns more similarly than they did abstract nouns. It was suggested above that the within-subjects reliability may have resulted from the fact that concrete nouns are better cues than abstract nouns. Hence, for concrete nouns, the subjects may have been better able to remember their previous definitions. This explanation cannot be used to explain why there was greater between-subject reliability in the definitions. Because concrete nouns represented a specific class of physical objects, the subjects may have agreed on the attributes of these objects to a greater extent than they agreed on the attributes of abstract concepts.

The results of this experiment show a modest reliability of retrieval from semantic memory with a 1-week test-retest period. Because semantic memory is considered to be memory in which often-used information is permanently stored, it was expected that repeated constructions of propositions describing a familiar noun would be similar in content, although retrieved from semantic memory at different times. In general, however, the percentage overlap between the two retrievals was approximately 50%. For the reliability of content measured here, it is difficult

to state what value would constitute a "good" reliability. Nevertheless, the result that only approximately half the information retrieved from semantic memory was retrievable 1 week later indicates that retrieval of information from semantic memory is a markedly probabilistic process. The contents of semantic memory may be quite stable over time periods of 1 week or so, but the accessibility of this information fluctuates to a surprising degree.

REFERENCES

ANDERSON, J. R. (1981). Concepts, propositions, and schemata: What are the cognitive units? In H. E. Howe, Jr., & J. H. Flowers (Ed.), *Nebraska Symposium on Motivation* (Vol. 28). Lincoln: University of Nebraska Press.

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

BELLEZZA, F. S. (1984). Reliability of retrieval from semantic memory: Common categories. *Bulletin of the Psychonomic Society*, **12**, 324-326.

CICCHETTI, D. V. (1972). Extension of multiple-range tests to interaction tables in the analysis of variance: A rapid approximate solution. *Psychological Bulletin*, **77**, 405-408.

CLARK, H. H., & CLARK, E. V. (1977). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

KINTSCH, W. (1972). Notes on the structure of semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory.* New York: Academic Press.

KIRK, R. E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Wadsworth.

LINDSAY, P. H., & NORMAN, D. A. (1977). *Human information processing.* New York: Academic Press.

MOORE, J., & NEWELL, A. (1973). How can MERLIN understand? In L. W. Gregg (Ed.), *Knowledge and cognition.* Potomoc, MD: Erlbaum.

OGDEN, C. K., & RICHARDS, I. A. (1923). *The meaning of meaning.* London: Routledge & Kegan Paul.

PAIVIO, A. (1971). *Imagery and verbal processes.* New York: Holt:

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76**(1, Pt. 2).

ROSCH, E., MERVIS, C. B., GRAY, W., JOHNSON, D., & BOYES-BRAEM, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words.* New York: Columbia University, Teacher's College, Bureau of Publications.

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

WITTGENSTEIN, L. (1968). *Philosophical investigations.* (G. E. M. Anscombe, Ed. and trans., 3rd ed.). New York: Macmillan.

(Manuscript received for publication April 4, 1984.)