Behavior Research Methods, Instruments, & Computers
2004, 36 (3), 421-431

Semantic distance norms computed from an electronic dictionary (WordNet)

WILLIAM S. MAKI, LAUREN N. McKINLEY, and AMBER G. THOMPSON
Texas Tech University, Lubbock, Texas

WordNet, an electronic dictionary (or lexical database), is a valuable resource for computational and cognitive scientists. Recent work on the computing of semantic distances among nodes (synsets) in WordNet has made it possible to build a large database of semantic distances for use in selecting word pairs for psychological research. The database now contains nearly 50,000 pairs of words that have values for semantic distance, associative strength, and similarity based on co-occurrence. Semantic distance was found to correlate weakly with these other measures but to correlate more strongly with another measure of semantic relatedness, featural similarity. Hierarchical clustering analysis suggested that the knowledge structure underlying semantic distance is similar in gross form to that underlying featural similarity. In experiments in which semantic similarity ratings were used, human participants were able to discriminate semantic distance. Thus, semantic distance as derived from WordNet appears distinct from other measures of word pair relatedness and is psychologically functional. This database may be downloaded from www.psychonomic.org/archive/.

The modern study of word association and the development of association norms was begun in the 1880s by Galton, Trautscholdt, and Cattell (Esper, 1973; Woodworth, 1938). Trautscholdt, working in Wundt's Leipzig laboratory, distinguished between two kinds of associations: *Outer* associations were formed by repetition and contiguity; *inner* associations were based on "semantic or logical relationships, among which were listed superordination, subordination, coordination, and causality" (Esper, 1973, p. 98). Esper deemed the distinction between inner and outer associations "unfortunate," perhaps because it could be taken to imply a nonempirical basis for semantic relations. But that aside, Trautscholdt's classification anticipated current research and theory on associative and semantic influences on priming (Lucas, 2000). About 20 years after Trautscholdt's work, Wreschner suggested that "a lexicon of associatively connected words could be the basis for many interesting studies, especially the psychology of language" (Esper, 1973, p. 97). These two ideas lay the groundwork for the research to be reported in this article—that word associations and semantic relations are separable and that large-scale databases have much to contribute to the understanding of associative and semantic relations between words.

Most of the subsequent work on normative information about relations among words followed in the tradition of associationism, focusing on word associations. The most ambitious project was begun in 1973 by Nelson and McEvoy and spanned 30 years. This effort resulted in the most comprehensive set of association norms ever produced (Nelson, McEvoy, & Schreiber, 1998). Their word association database contains 5,019 normed words and 72,176 responses (i.e., associative values for 72,176 pairs). These norms were collected using the method of free association, in which participants are instructed to respond to a stimulus word with the first word that comes to mind. The associative values (strengths) are then obtained by computing the relative response frequencies (i.e., probabilities that a particular response will be evoked by a given stimulus).

What are the origins of word associations? One answer, again following in the associationist tradition, is that word associations arise from repetition and contiguity as manifested in the frequency with which the words co-occur (e.g., Fischler, 1977). Lexical co-occurrence is correlated with associative strength (Spence & Owens, 1990) and has been proposed as a less costly and more reliable source of association norms (Church & Hanks, 1989). The advent of very large corpora of computer-readable text has made it possible to build large semantic spaces derived from co-occurrence statistics (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996).

In contrast to the availability of association and co-occurrence databases, normative information about semantic relations has been slow to develop. There has been a tendency to select first and norm second (e.g., Fischler, 1977) and then for subsequent investigations to rely on those materials (e.g., Fischler, 1977, via Seidenberg, Waters, Sanders, & Langer, 1984, to Thompson-Schill, Kurtz,

The database, in the form of a comma-separated-value file (usfjcnlsa.csv) is available for downloading from ftp://ftp.ttu.edu/pub/maki. We are deeply indebted to several people for their generous sharing of their own computational resources, experimental results, and technical information; in alphabetical order, they are Tom Landauer, Ken McRae, Doug Nelson, Ted Pedersen, José Quesada, Mark Steyvers, and Randi Tengee. We thank Doug Nelson and two anonymous reviewers for insightful comments on an earlier version of this article. Correspondence concerning this article should be addressed to W. Maki, Department of Psychology, Texas Tech University, Lubbock, TX 79409 (e-mail: bill.maki@ttu.edu).

& Gabrieli, 1998). McRae, de Sa, and Seidenberg (1997) highlighted the problems inherent in relying on intuition for the selection of semantically related pairs of words. To be sure, several sets of semantic norms have been obtained since (see the review by McRae, Cree, Seidenberg, & McNorgan, in press). But these norms are not readily available and consist of a few dozen to, at most, a couple of hundred words. The lack of semantic norms is an impediment to research on semantic relationships; notably, the absence of normative resources has made it difficult to sort out the contributions of associative and semantic relations to priming effects (Lucas, 2000). Perhaps the development of norms for semantic relations has been delayed because of the high cost of obtaining them (as has been observed by Budanitsky & Hirst, 2000, and Church & Hanks, 1989).

Some recent progress has been made toward developing a more comprehensive set of semantic norms by McRae and colleagues (Cree & McRae, 2003; McRae et al., in press; McRae et al., 1997). McRae used a feature-norming procedure in which participants listed features of a word's referent, including physical, functional, and categorical properties, as well as encyclopedic facts. The result, for each word, is a vector of frequencies for each of the many features present in the norms. Featural similarity between words can then be determined by computing the cosine between the vectors for each pair of words. Feature norms are important for investigating claims about semantic representation made by distributed memory models (such as in Moss, Hare, Day, & Tyler, 1994). However, even after 12 years of norming efforts, feature norms are available for only 541 words (McRae et al., in press). So semantic resources still lag way behind the thousands of words present in associative and co-occurrence databases.

In this article, we propose another source of semantic information that may greatly enhance our ability to select large numbers of pairs of words, a priori, on the basis of semantic relations. We have used an electronic dictionary, WordNet (Fellbaum, 1998), to build a relatively large scale database of semantic distances. We have combined these semantic distances with existing co-occurrence, semantic, and associative norms (Landauer & Dumais, 1997; McRae et al., in press; Nelson et al., 1998). The resulting database has been successfully used to select stimuli for experiments on similarity ratings and memory. In the remainder of this article, we will describe the methods and materials used in constructing our semantic distance database. Then we will highlight some results pertinent to our claims about the utility of this resource and its psychological properties.

METHODS AND MATERIALS

WordNet

The electronic dictionary, WordNet, was conceived in the 1980s by George Miller. Miller has described the motivation for and history of the development of WordNet in the forward to a book of chapters describing its structure and uses (Fellbaum, 1998; see also Miller, 1999). However, in the same book, Miller (1998) also observed that, although WordNet has been popular among computational linguists, it has been mostly ignored by psychologists. The WordNet Web site contains a link to a bibliography (http://www.cogsci.princeton.edu/~wn/papers.shtml) that lists over 300 articles germane to WordNet. In contrast, an electronic search of the PsycINFO database for "WordNet" returned only five articles—a chapter by Miller, a sociological article, and three papers on applications of artificial intelligence. The bibliography lists numerous papers using WordNet for studies of word sense disambiguation, but WordNet apparently has had no impact on psychological studies of disambiguation (or anything else). One possible and simple reason for this neglect may be that it has not been obvious how to extract (potentially) useful information from WordNet. Only recently have computational tools been made available that allow WordNet to be mined for semantic distances (Patwardhan & Pedersen, 2003).

WordNet is a lexical database consisting of (at this writing) 115,424 nodes. Each node contains one or more synonymous words (*synsets*). The largest group of nodes (69%) contain nouns; the remaining synsets contain adjectives (16%), verbs (12%), and adverbs (3%). The noun synsets are linked by relationships, notably hypernymy–hyponymy (super- and subordinates; *is-a*) and meronymy–holonymy (*has-a* and *is-part-of*). Figure 1 shows a part of the WordNet taxonomy linking *mouse*, *rat*, and *cat*. *Mouse* and *rat* are both rodents; the feline *cat* is a carnivore. The more general taxonomy, beginning with *placental mammal*, is common to both *rodent* and *carnivore* categories. The root node, *entity*, subsumes all other things in the taxonomy.

Different Approaches to Semantic Distance

The structure in Figure 1 squares with our intuitions that *mouse* and *rat* are more similar than are *mouse* and *cat* (even though in the association norms, *mouse* and *cat* are more highly associated). The question is how to measure that similarity. Several measures were proposed and investigated by computational linguists in the late 1990s. There are three kinds of such measures (Jiang & Conrath, 1997). The edge-based approach (e.g., Leacock & Chodorow, 1998) computes some transform of the number of edges (links) between concepts in a graph such as that in Figure 1. The node-based approach (Resnik, 1995) begins with the observation that the information content of a concept depends on its location within the graph. More precisely, the lower the probability of encountering a concept, the higher is its information value. The concept *entity* is associated with all things, so its information value is zero. The concept *rodent* is infrequently encountered in the taxonomy, so its information value is higher (than *entity*, higher than *mammal*, and higher than *placental mammal*). In this approach, frequency is estimated from available frequency counts of

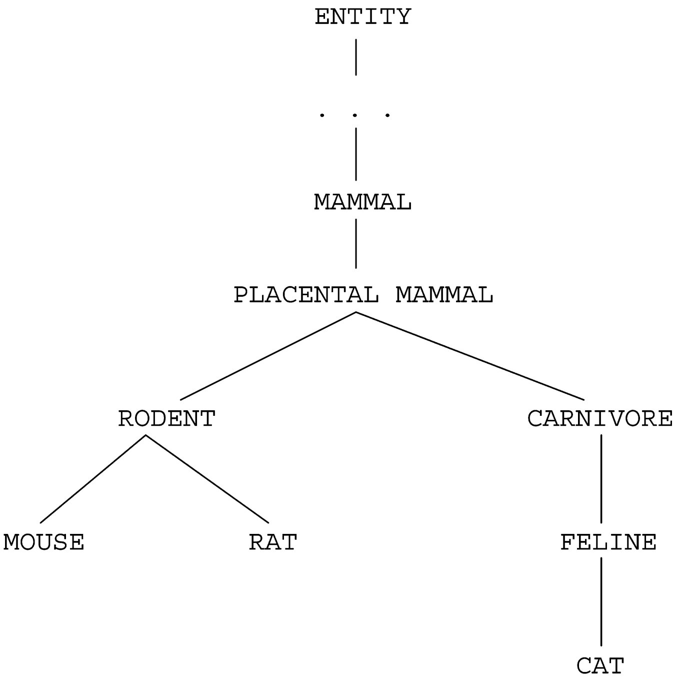

Figure 1. A portion of the WordNet taxonomy showing hypernym and hyponym (*is-a*) relations for *mouse*, *rat*, and *cat*.

words in English text (such as the Brown corpus; Francis & Kučera, 1982). These frequencies are then used to compute the probability for each concept.[1] The information content (IC) for concept c is quantified as $\mathrm{IC}(c) = -\log p(c)$. Because *entity* occurs in each pathway in the taxonomy, its probability is 1, and its IC is 0. *Mammal*'s IC would be considerably higher, and *rodent*'s IC higher still. The node-based measure of similarity between two concepts, c_1 and c_2, is the maximum IC of all the nodes subsuming c_1 and c_2 in the taxonomy.

The third approach is the one developed and investigated by Jiang and Conrath (1997) and involves a combination of edge- and node-based approaches. Three IC values are required to compute the Jiang and Conrath measure of semantic distance between two concepts, c_1 and c_2. One needs the IC values for c_1 and c_2 and also the IC value for the lowest superordinate of both c_1 and c_2, c_0. The distance (here abbreviated as JCN) is given by $\mathrm{JCN} = \mathrm{IC}(c_1) - \mathrm{IC}(c_0) + \mathrm{IC}(c_2) - \mathrm{IC}(c_0)$. This is the sum of the informational difference between each of the two concepts and their lowest superordinate.

Which of these approaches should be implemented? There are two kinds of evidence, one psychological and one computational, that combine to indicate that the node-based approach is superior to the edge-based approach and that, in turn, the combined measure (JCN) is superior to the node-based approach. Miller and Charles (1991) have provided a set of semantic similarity ratings for 30 pairs of nouns obtained from human observers; this data set has been used as the benchmark for assessing the various computational measures of semantic distance. Resnik (1995) compared the node-based distance measures against the Miller and Charles ratings. Resnik replicated the Miller and Charles study with 10 human subjects. He reported that the correlation with the original Miller and Charles data was $r = .90$. The correlation of the edge-based measure of similarity was $r = .66$, but the correlation of the node-based measure was $r = .79$. It appears from these results that the node-based information content measure is a better predictor of the human data.[2] Jiang and Conrath (1997) correlated edge-based, node-based, and their own combination measures with the Miller and Charles ratings. Consistent with Resnik, they reported that the node-based measure correlated more highly ($r = .819$) than did the edge-based measure ($r = .604$); however, the combination measure (expressed as semantic similarity) correlated more highly yet ($r = .865$). (Later in this article, we will report additional evidence supporting the superiority of the JCN measure.)

The second kind of evidence bearing on the "best" measure of semantic distance comes from a study of automatic spell checking by Budanitsky and Hirst (2001). They were interested in a problem that afflicts the current generation of spell checkers (such as those implemented in word-processing systems). Spell checkers do not flag malapropisms—properly spelled words that appear in the wrong contexts. For example, in an essay on dairy farming, misspelling dairy as diary would be a malapropism. These errors are ubiquitous and, hence, a significant problem for automatic detection. The belief that our national defense should include a strong detergent has been attributed to a popular television comedian. A front-page article in a newspaper reported that "there was some descent among current and former regents." And a politician has been reported to mention, for example, the reduction of "greenhouse gas admissions." Budanitsky and Hirst compared five measures of semantic distance computed from WordNet (including the three mentioned above) with respect to their ability to (accurately) nominate words as malapropisms. The combination (JCN) measure was significantly more accurate at detecting malapropisms than any of the other measures. Thus, the available evidence combines to nominate the JCN measure as the one with the most potential as a measure of semantic distance in WordNet.

Computing Semantic Distance for a Database

The various measures studied by Budanitsky and Hirst (2001) have been programmed by Patwardhan and Pedersen (WordNet-Similarity; 2003) and distributed as a collection of Perl modules. (See Schwartz, 1998, for an introduction to the Perl language.[3]) These modules in turn make use of QueryData, a Perl module for accessing WordNet data (Rennie, 2000). The computations included in the semantic distance database reported here were performed at the Texas Tech University High Performance Computing Center, using an SGI Origin 2000 with 56 nodes. Each node has a MIPS R1200 processor running at 300 MHz. The time in seconds (t) taken by WordNet-Similarity-0.05 to compute N distances was found to be estimated by $t = 30.63 + 0.373\ N$. Computing 63,000 distances took 37 CPU hours; computing distances for all unique pairings of 4,150 words took 37 CPU *days*. The actual computational time was shortened considerably by parallel processing of sublists of pairings. Only the JCN measure was computed. The Brown corpus (Francis & Kučera, 1982) was used to estimate frequencies.[4] At the time the computations were performed (summer, 2003), the current version of WordNet was 1.7.1.

The semantic distance database reported here contains semantic distances for 49,559 noun–noun and verb–verb pairs. These pairings are based on 4,150 nouns and verbs that appear in the Nelson et al. (1998) word association norms. The database only includes pairs for which semantic distance could be computed. The JCN values in the database average 11.5, with a standard deviation of 6.6. The median JCN value is 12.3; the values range from 0 to 30.5.[5]

The semantic distance database was motivated by the desire to select word pairs on the basis of semantic or associative relations while measuring (if not also controlling) other aspects of word relatedness. Thus, the database also contains several measures in addition to the JCN values. Forward and backward associative strengths were obtained from the Nelson et al. (1998) association norms. These strengths are the probabilities of a word's being given as a response to a cue word in free association tests. For example, the probability that *mouse* will elicit a response of *cat* is .543, but the probability that *cat* will elicit *mouse* is .256. The forward strength (FSG) for *mouse–cat* is .543, and the backward strength (BSG) for that pair is .256. We included word frequencies (based on the Brown corpus, as supplied by Nelson et al., 1998). Similarity values (cosines) based on latent semantic analyses (LSAs; Landauer & Dumais, 1997) provide another measure of semantic organization based on lexical co-occurrence. The LSA cosines included in our database were computed from a large word × document co-occurrence matrix (the "General Reading up to 1st Year College" corpus).[6]

SOME CHARACTERISTICS OF SEMANTIC DISTANCE

In the remainder of this article, we will report on a series of studies that bear on the utility of using semantic distance and on two important questions. (1) Is semantic distance, as computed from a dictionary, psychologically functional? (2) If so, is it distinct from other measures of word relatedness?

Correlations Between Distance and Other Measures

We computed several correlations between the JCN distance and other measures previously reported in the computational linguistics literature on semantic distance. In the first set, the present computations were compared with those of Resnik (1995) and Jiang and Conrath (1997). The correlations of the JCN distance measure[7] with the human similarity ratings obtained by Miller and Charles (1991) and by Resnik were $r = -.876$ ($N = 30$) and $r = -.891$ ($N = 28$), respectively. (Two of the pairs were not included in Resnik's analyses.) These pairs of words were selected by Miller and Charles from a larger set of 65 pairs originally created by Rubinstein and Goodenough (1965). The correlation between the JCN distance measure and the full set of ratings obtained by Rubenstein and Goodenough was $r = -.842$ ($N = 65$). The correlation with the original similarity measure reported by Jiang and Conrath was $r = -.935$ ($N = 30$). That correlation is not perfect because of variations in the version of WordNet and the corpus used to estimate frequency; nevertheless, the high correlations show that the computations reported here correspond nicely to those previously reported. (The correlations are all significant at better than $p < .05$, the significance to be used in reporting results throughout the remainder of this article.)

If distance measures contain information different from co-occurrence statistics (Niwa & Nitta, 1998), we should not expect particularly high correlations between the distance and the LSA measures, because they may measure different aspects of semantic relatedness. Using all 49,559 pairs from the full database, we computed the correlations between the JCN distance measure, (forward) associative strength, and the LSA cosine measure. The correlations between the JCN and the other two measures were $r = .146$ ($N = 49{,}559$) and $r = -.158$ ($N = 49{,}362$) for the strength and the LSA values, respectively. The correlation between the strength and the LSA measures was $r = .267$ ($N = 49{,}362$). Although these correlations are all highly significant, their magnitudes indicate a substantial amount (over 90%) of unshared variance among the three measures.

Collins and Loftus (1975) have distinguished between semantic distance (as the shortest path between nodes in a semantic network) and semantic relatedness (as the aggregate of all paths). Thus, it is possible that the two measures for which there are now substantial norms may not be measuring the same thing. The JCN distance measure has some of the flavor of a pathway measure (as indeed it should, being a composite of edge- and node-based approaches). The featural similarity measure based on semantic feature norms (McRae et al., in press) has no such pathway component. Thus, the relationship of semantic distance and featural similarity is an interesting empirical and theoretical question. We created 132,355 pairs of words based on 515 of the 541 concepts in the McRae et al. (in press) feature norms (25 homographs were excluded from the analysis, and one concept was not found in WordNet). For each pair of words, the JCN values were computed, and feature vector cosines were obtained.[8] The correlation was $r = -.345$ ($N = 132{,}355$). Although statistically significant, that correlation still leaves about 88% unshared variance.

The conclusion we draw from these correlations is that the measures considered here are intercorrelated, but weakly so. The strongest correlation was between semantic distance and semantic featural similarity. Although even that relationship left considerable variance unexplained, the correlation is still over twice those found between JCN and the strength and co-occurrence measures. Apparently, as we will confirm next, semantic distance and featural similarity both reflect some of the same aspects of semantic organization.

Hierarchical Cluster Analyses of Knowledge Types

McRae et al. (in press) obtained frequencies with which human raters assigned semantic features to 541 words. Cree and McRae (2003) used these feature norms to study the structure of 34 common categories (such as *mammal*, *fish*, *automobile*, *tool*, *weapon*, *vegetable*, and *fruit*). Each category was represented as a frequency distribution in which each entry corresponded to the frequency with which the exemplars of that category exhibited one of 22 feature types. The resulting 34 × 22 matrix was subjected to a hierarchical cluster analysis (see below). The resulting dendrogram revealed three main subclusters—creatures, nonliving things, and fruits/vegetables. Fruits/vegetables clustered with nonliving things prior to being linked at the top level to the creature cluster. This structure is interesting because it is congruent with trends identified by Cree and McRae in the neuropsychological literature concerning category-specific semantic deficits.

The question we ask here is whether the JCN semantic distance measure will exhibit similar structural properties. We computed JCN measures for all pairs of 31 of the 34 categories studies by Cree and McRae (2003). The three categories that contained more than one term were dropped (e.g., *fashion accessory*). The *root*/*tuber* category was represented as just *tuber*. These values were cast as a symmetrical 31 × 31 matrix. Thus, each category's row vector represented its similarity to and dissimilarity from each of the other categories. Otherwise, the clustering method was performed exactly like that reported by Cree and McRae. SPSS was used to perform the average-linkage between-groups hierarchical cluster analysis. The cosine for each pair of category vectors was used as the measure of similarity. The resulting dendrogram is shown in Figure 2.

Like the dendrograms reported by Cree and McRae (2003, Figures 2 and 4), Figure 2 shows three clusters of creatures, nonliving things, and fruits/vegetables. Also like the Cree and McRae results, nonliving things and fruits/vegetables cluster at a high level. However, at a finer grain, the dendrograms show substantial differences. For example, *gun* and *weapon* clustered together before being combined with *tool* and *utensil*, which clustered together. In Cree and McRae's results, *weapon* and *clothing* clustered together before being combined with *gun*; *tool* clustered with *machine*, and *utensil* clustered with *container*. We are not sure of the reasons for these differences. But a coarse-grain/fine-grain view of the dendrograms suggests a correspondence with the correlational results. At the top level, semantic distance and featural similarity measures appear to measure the same things, but at a finer grain level, there are many differences (showing up in the correlational analyses as unshared variance).

Experiments on Human Ratings of Semantic Distance

The pairs of nouns studied by Miller and Charles (1991), by Resnik (1995), and then by Jiang and Conrath (1997) were specially selected by Rubinstein and Goodenough (1965) so as to include both synonymous and nonsynonymous pairs. Inspection of the pairs reveals that the nonsynonymous pairs were created by explicitly unpairing words from the synonymous pairs. That raised a concern that the correlations reported so far might be forced by the specific pairs created by Rubinstein and Goodenough. Consequently we conducted our own rating studies to confirm that semantic distance extracted from WordNet could be discriminated by human observers. We performed two experiments that were nearly identical, except for the category boundaries and range of JCN values used to select pairs.

Method. Each experiment included 225 pairs of words from the semantic distance database. In each experiment, the pairs were selected on the basis of both forward associative strength (FSG) and semantic distance (JCN). Strengths were classified as high (FSG $\geq .6$), medium ($.35 \leq$ FSG $\leq .5$), or low ($.10 \leq$ FSG $\leq .25$) in both experiments. Backward strengths were minimized ($<.10$). In Experiment 1, upper and lower boundaries of five se-

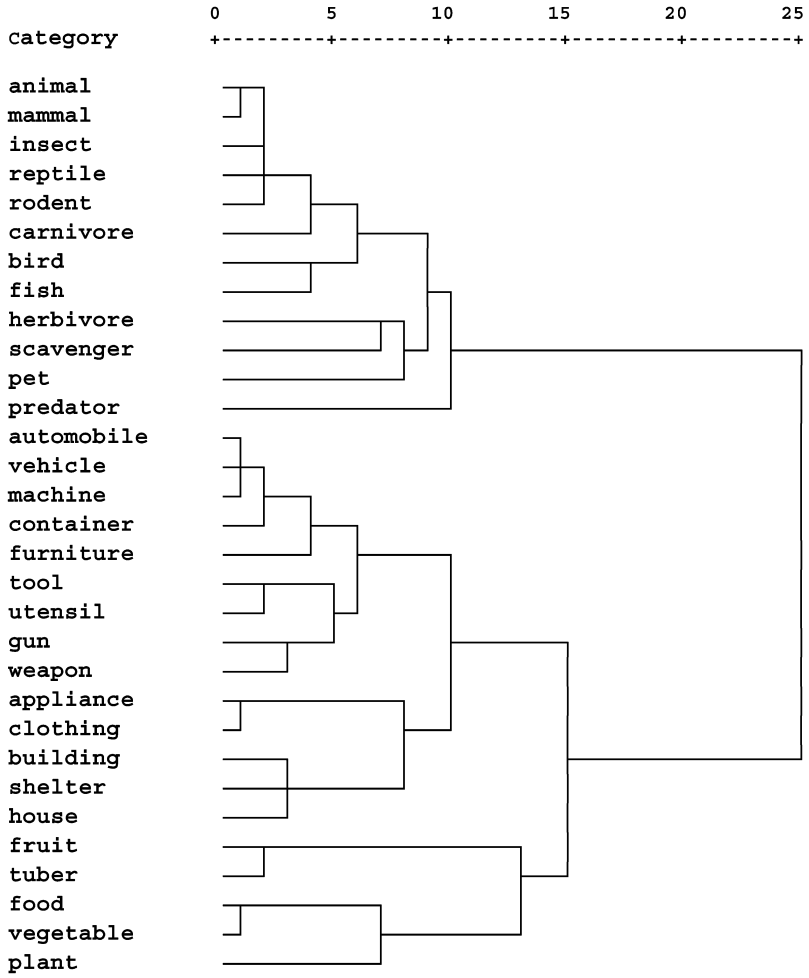

Figure 2. Dendrogram from hierarchical cluster analysis based on semantic distances between categories (cf. Figure 2 in Cree & McRae, 2003).

mantic distance categories were separated by five JCN units (0–5, 5–10, . . . > 20). In Experiment 2, the boundaries were separated by four JCN units (0–4, 4–8, . . . 16–20). In each experiment, 15 pairs were randomly selected from each of the 15 combinations of strength (3) × distance (5). The averages and ranges for strengths and distances within each of the strength–distance combinations are presented as tables in the Appendix.

The participants were instructed to rate the pairs on the basis of similarity of meanings. The pairs of words were presented one at a time, with the two words in each pair appearing side by side on a computer screen. A 7-point scale appeared below the words. A rating of 1 on the scale meant that *the words were completely different in meaning*; a rating of 7 meant that *the words were identical in meaning*. The participants responded by typing their numerical ratings on the computer keyboard.

The pairs of words were presented in a different random order for each participant in the experiments. The participants were recruited from the Texas Tech University psychology participant pool and were compensated for course credit. $N = 42$ for Experiment 1, and $N = 45$ for Experiment 2.

Results. In both experiments, human ratings of similarity declined with increasing distance. Also, in both experiments, similarity ratings were affected by associative strength, with higher ratings given to more strongly associated pairs. But the two variables had independent influences on ratings; the interactions were not reliable. Figure 3 shows the mean similarity ratings for each combination of distance and strength in each experiment. These plots show the effects of computed distance on similarity ratings and provide evidence for our further contention that the associative effect was due mainly to the high ratings given to highly associated pairs.

Analyses of variance (using interactions with items as error terms) showed that the effects of distance were reliable in both Experiments 1 and 2 [$F(4,210) = 18.73$ and $F(4,210) = 19.03$, respectively]. The effects of associative strength were also reliable in both experiments

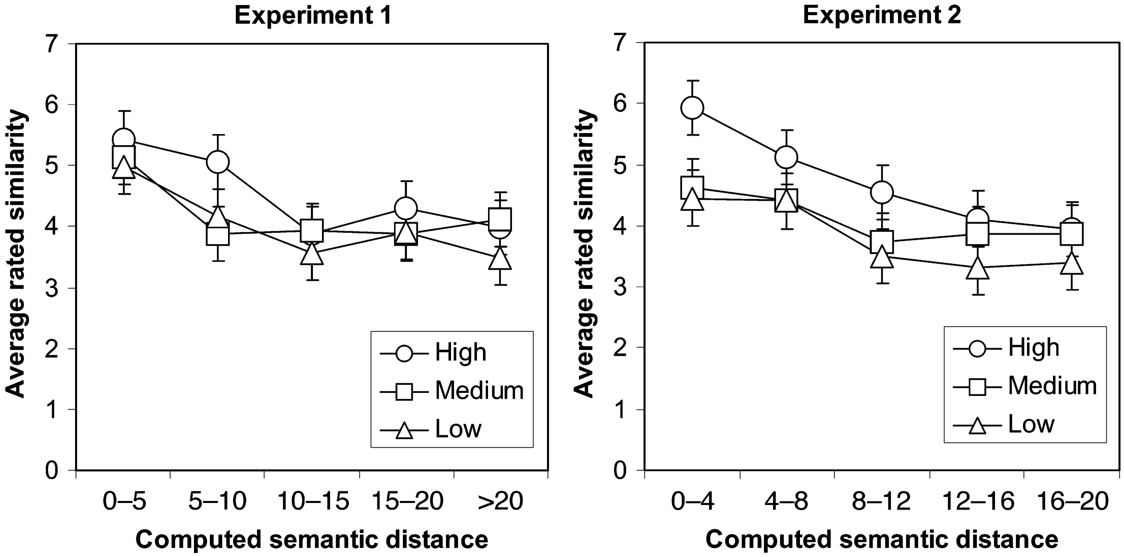

Figure 3. Average rated similarity as a function of computed semantic distance and associative strength. Strengths were selected to have high, medium, or low associative values; see text for the details. Confidence limits of 95% are shown for each mean.

[F(2,210) = 6.41 and F(2,210) = 20.90]. In both experiments, Tukey HSD tests showed that the high-strength pairs were rated significantly higher than either the medium- or the low-strength pairs, which did not differ significantly. Neither of the interactions between distance and strength was significant [F(8,210) = 1.44 and F(8,210) = 1.31].

The effects of semantic distance, averaged across associative strengths, are shown in Figure 4. Similarity ratings from the human observers have been transformed into distance ratings by reverse scoring. Also shown are best-fitting linear regression plots for low- and high-distance pairs. In both experiments, ratings linearly increase for the low-distance pairs and then show little variation for the high-distance pairs. The two linear plots intersect at JCN = 11.9, a point near the median of all the JCN values in the database (12.3).

Although we selected word pairs categorically, the actual JCN and FSG values appeared nearly continuous (see the ranges in the Appendix). Thus, we conducted additional multiple regression and correlation analyses. Because the selection of word pairs was done independently in each experiment, 48 pairs appeared in both experiments. Mean ratings were computed for each pair (as in the item analyses above) for each experiment. The correlation between experiments (r = .95, N = 48) indicated a high degree of reliability in the ratings.

The ratings for the 402 unique pairs were then averaged across experiments and entered into multiple regression analyses. These analyses included the two main variables of interest (JCN and FSG), as well as their interaction. Also included were two other strength variables from the Nelson et al. (1998) norms—BSG and MSG (mediated strength, a measure of strength of indirect associations)—and LSA. A simultaneous multiple regression showed that only the JCN and FSG variables were significant predictors of rated similarity (β = −.331 and .331, respectively). A stepwise multiple regression showed that JCN and FSG accounted for 24.7% of the variance in ratings (18.1% and 6.6%, respectively); the remaining variables accounted for only 0.8% of the variance. These results provide additional statistical support for the contention that associative strength and semantic distance were main (and independent) contributors to rated semantic similarity in these experiments.

One other set of analyses were performed on the mean ratings from the 402 pairs. We used the WordNet-Similarity software (Patwardhan & Pedersen, 2003) to compute the edge-based (Leacock & Chodorow, 1998) and node-based (Resnik, 1995) measures of semantic relatedness among the pairs. The simultaneous multiple regression showed that only the JCN measure was a significant predictor of rated similarity (β = −.292). A subsequent stepwise regression showed that, although the JCN measure accounted for 18.1% of the variance in ratings, the other computed measures accounted for only 0.1%. This is converging evidence supporting our earlier nomination of JCN as the best available measure of semantic distance within WordNet (see also Budanitsky & Hirst, 2001; Jiang & Conrath, 1997).

Discussion. These experiments help answer the two questions we asked about computed semantic distances. First, these experiments show that semantic distance is psychologically functional, in that humans are sensitive to semantic distance computed from WordNet. Second, these experiments suggest that semantic distance and as-

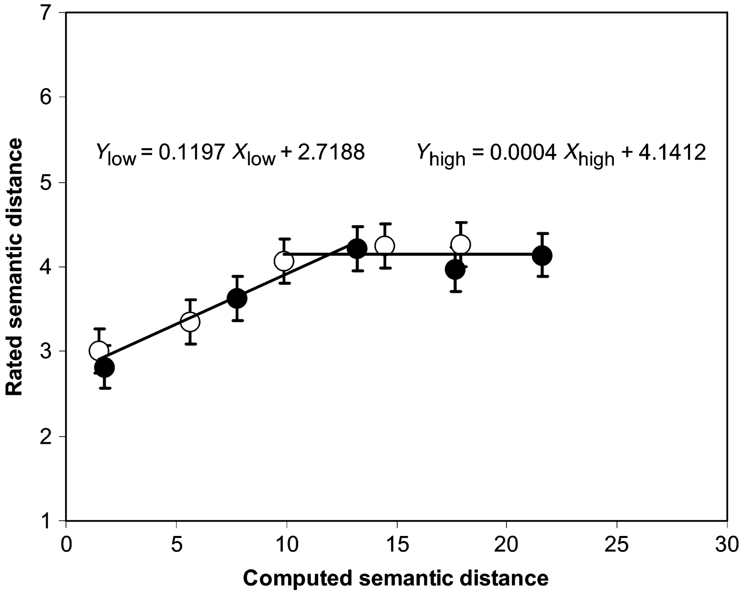

Figure 4. Mean rated semantic distance as a function of computed semantic distance. The semantic distance was computed from WordNet 1.7.1, using the Jiang and Conrath (1997) combination method. Filled symbols represent average ratings obtained in Experiment 1; open symbols represent average ratings obtained in Experiment 2. Confidence intervals (95%) are included for each data point. The regression equations (and linear plots) are based on the first three (*low*) and last three (*high*) data points for each experiment (six points per plot). The plots intersect at a computed distance of 11.9.

sociative strength exert independent influences on semantic similarity judgments. However, the fact that strongly associated pairs were judged more semantically similar raises other questions that need exploration. Under what circumstances does associative strength influence semantic judgments? How general is the additivity observed in these experiments?

The experiments reported here also provide some guidance for researchers who would use the semantic distance norms for selection of stimulus materials. There appeared to be little influence of associative strength on semantic similarity ratings for low to medium strengths. Semantic distance was discriminable only for the lowest half of the distribution of JCN values. Thus, confining selection of materials to these ranges seems advisable.

CONCLUSIONS, IMPLICATIONS, AND FUTURE DIRECTIONS

In this article, we have reported on the need for, implementation of, and psychological properties of a database composed of semantic distances. The semantic distances were obtained from computations performed on an electronic dictionary, WordNet (Fellbaum, 1998). We identified a particularly promising measure of semantic distance (Jiang & Conrath, 1997) and used recently developed software (Patwardhan & Pedersen, 2003) to create the database of distances. Included in the database are other measures of word relatedness: associative strength (Nelson et al., 1998) and similarity based on LSA (Landauer & Dumais, 1997). We showed that it is possible to select pairs of words on the basis of the orthogonal combination of associative strength and semantic distance. Experiments on the rating of semantic similarity showed that human participants can discriminate semantic distances; semantic distance, then, although obtained computationally, is psychologically functional. Moreover, semantic distance is not well correlated with measures of associative strength and co-occurrence, which, along with the observed additive effects of strength and distance on human ratings, suggests that semantic distance is separable from other measures of semantic and associative relatedness.

The results of the regression analyses comparing various computational measures of semantic distance in WordNet add to the evidence favoring the Jiang and Conrath (1997) informational distance measure over purely edge- or node-based measures. Previous evidence included the superiority of their measure in detecting malapropisms in a spell-checking application (Budanitsky & Hirst, 2001) and the correlations between the computational measures and the human ratings reported by Jiang and Conrath. In support of Jiang and Conrath's correlations, we showed in the analyses of our own data that their measure was the best predictor of human ratings of semantic similarity. Exactly why this is so remains an open question (a question also raised by Budanitsky & Hirst, 2001).

Our combination of a semantic distance measure with preexisting word association statistics into a common database should support new lines of research and theory on associative and semantic memory. Armed with the semantic distance database, we have launched a program of research examining the relations between semantic distance, featural similarity, associative strength, and similarity based on LSAs. These studies confirm the conclusion drawn here—namely, that semantic distance does influence human judgments (and memory) and does so mostly independently of associative strength, lexical co-occurrence, and featural similarity.

Our work on semantic distance computed from WordNet may spur further cognitive psychological research in which WordNet is used as a resource. For example, it may be possible to automatically mine WordNet for those semantic features nominated by human observers (as in Cree & McRae, 2003) and, thus, greatly expand the corpus of words and categories for which semantic feature norms are available.

The semantic distance database may prove useful for the continuing effort to resolve theoretical issues concerning the associative and semantic aspects of word relatedness. We perceive two general lines of thought about the relations between word associations and semantics. One was articulated by Deese in his concept of associative meaning: "The distribution of responses evoked by a particular word as stimulus defines the meaning of that word" (Deese, 1965, p. 43). Thus, semantics are derived from word associations. The recent derivation of word association spaces by Steyvers, Shiffrin, and Nelson (in press) follows in that tradition; in their work, semantic similarity between pairs of words was quantified by computing the distance along the shortest path between those words through a network of associative links based on free association norms (Nelson et al., 1998).

The other line of thought is traceable to Trautscholdt's distinction between associative and semantic relations (Esper, 1973). On this view, co-occurrence gives rise to separate associative and semantic relations (perhaps corresponding to the lexical and semantic networks conceived by Collins & Loftus, 1975). Garskof and Forrester (1966) obtained ratings of both associative strength and semantic similarity and found that there was a low correlation between the two; they concluded that association strength and semantic similarity are "distinct types of word relatedness." More recently, Lund, Burgess, and Audet (1996) likened the distinction between associative and semantic information to the distinction between local and global co-occurrence, where the latter is computed from dimensional reductions, as in their hyperspace analogue to language model or in LSA (Landauer & Dumais, 1997). This separation of associative and semantic information suggests that measurements based on word associations are likely to be only weakly (if at all) related to measures of semantic relatedness, such as distance (as in this article) or feature overlap (as in Cree & McRae, 2003). That suggestion is consistent with the low correlation we found between semantic distance and associative strength.

The theoretical picture is likely to become further complicated by another of our observations. We computed semantic distances on the basis of an electronic dictionary and found that people were sensitive, in their ratings of semantic similarity, to those distances. Cree and McCrae (2003) computed cosines between pairs of feature vectors. The cosine measures the distance in hyperspace between vectors. Yet we found that these two measures of distance were only weakly correlated. One conclusion from this observation is that these two measures are tapping different kinds of semantic information. These observations reinforce one of our opening conjectures: Large-scale databases of the sort presented here will help refine our understanding of associative and semantic information in human memory and cognition.

REFERENCES

Budanitsky, A., & Hirst, G. (2001). *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. Paper presented at the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (pp. 76-83). Vancouver, BC: Association for Computational Linguistics.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407-428.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, and cello (and many other such concrete nouns*). Journal of Experimental Psychology: General*, **132**, 163-201.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore: Johns Hopkins University Press.

Esper, E. A. (1973). *Analogy and association in linguistics and psychology*. Athens: University of Georgia Press.

Fellbaum, C. (Ed.) (1998). WordNet: An electronic lexical Database. Cambridge, MA: MIT Press. Available at http://www.cogsci.princeton.edu/~wn.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, **5**, 335-339.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Garskof, B. E., & Forrester, W. (1966). The relationships between judged similarity, judged association, and normative association. *Psychonomic Science*, **6**, 503-504.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265-283). Cambridge, MA: MIT Press.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, **7**, 618-630.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**, 203-208.

Lund, K., Burgess, C., & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 603-608). Mahwah, NJ: Erlbaum.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (in press). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, & Computers*.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, **126**, 99-130.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 23-46). Cambridge, MA: MIT Press.

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology*, **50**, 1-19.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language & Cognitive Processes*, **6**, 1-28.

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, **6**, 413-427.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Available at http://w3.usf.edu/FreeAssociation.

Niwa, Y., & Nitta. Y. (1998). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 304-309). San Francisco: Morgan Kaufman.

Patwardhan, S., & Pedersen, T. (2003). *WordNet::Similarity*. Available at http://search.cpan.org/dist/WordNet-Similarity/.

Rennie, J. (2000). *WordNet::QueryData: A Perl module for accessing the WordNet database*. Available at http://www.ai.mit.edu/people/jrennie/WordNet/. (Also available from http://www.cpan.org/)

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 448-453). San Francisco: Morgan Kaufman.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**, 627-633.

Schwartz, A. (1998). Tutorial: Perl, a psychologically efficient reformatting language. *Behavior Research Methods, Instruments, & Computers*, **30**, 605-609.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. *Memory & Cognition*, **12**, 315-328.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, **19**, 317-330.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (in press). Word association spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory & Language*, **38**, 440-458.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

NOTES

1. As in Resnik (1995), the frequency of each concept in the taxonomy is the sum of the frequency of that concept and the frequencies of all its children. For example, each occurrence of *mouse* counts toward the frequency of *mouse* and all subsuming concepts in the taxonomy (*rodent*, *placental mammal*, . . . , *entity*). Probabilities are based on the total frequency (of the root concept, *entity*). Information content files based on frequency counts from various corpora, including the Brown corpus, are available at http://www.d.umn.edu/~tpederse/similarity.html.

2. Although Resnik (1995) reported that the information content measure was "significantly better" than edge counting, the significance level was not reported.

3. The Perl language is available for download without charge from http://www.ActiveState.com and is routinely distributed with the Unix operating system.

4. The Brown corpus is based on 1,014,232 words contained in fiction and nonfiction texts printed in 1961 in the United States. An anonymous reviewer of the previous version of this article questioned the use of the Brown corpus, given its age. Moreover, it might be argued that the Brown corpus is small relative to newer corpora, such as the British National Corpus (BNC; http://www.natcorp.ox.ac.uk) that contains over 100 million words. Most of the BNC text is more recent than the Brown corpus, with about 90% of the text printed during 1975–1993. So do the age and/or size of the corpus matter? We recomputed JCN values for all the word pairs in our database, using the BNC information content file (see note 1). The correlation between $\text{JCN}_{\text{Brown}}$ and JCN_{BNC} was $r = .994$ ($N = 49{,}559$). Apparently, the corpus used to estimate frequency (and thus information content) does not play a large role in computing the JCN measure of semantic distance.

5. The file is not 100% complete. Many instances of plural nouns do not appear because the semantic distance computation returns the base form of the nouns; this should be corrected in a future version of the file.

6. The LSA computations were performed by Jose Quesada at the University of Colorado. The LSA cosines correspond only approximately to those that can be computed on the LSA Web site. The values used here are based on 419 factors; the values computed on the Web site are based on 300 factors.

7. Semantic distance is the complement (or inverse) of semantic similarity. The correlations reported by Jiang and Conrath (1997) were based on the JCN measure expressed as similarity (maximum possible distance minus computed distance), so the correlations they report are positive. Our database contains the actual computed distances, so the correlations with measures of similarity are negative.

8. The feature norms, courtesy of Ken McRae, were obtained from http://amdrae.ssc.uwo.ca/downloads.html.

ARCHIVED MATERIALS

The following materials and links may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, http://www.psychonomic.org/archive/.

To access these files or links, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Maki), and the publication year (2004).

File: Maki-BRMIC-2004.zip.

Description: The compressed archive file contains two files:

usfjcnlsa.csv, containing the norms developed by Maki et al. (2004), as a 3.1-MB comma-delimited text file generated by Excel 2003 for the PC. Each row represents one of 49,559 pairs of words. Listed within each row are the cue and target words and nine dependent measures. Forward (FSG), backward (BSG), and mediated (MSG) associative strength taken were taken from the Nelson et al. (1998) word association norms, as were cue and target concreteness and cue and target frequency values. The last two variables are the semantic distance (JCN) and the latent semantic analysis (LSA cosine) measures.

notes_on_usfjcnlsa.txt, a full description of the content of usfjcnlsa.csv, including definitions of the columns of the document and links to previous norms and related papers (a 5K plain text file).

The database and notes files, usfjcnlsa.csv and notes_on_usf_jcn_lsa.txt, also may be downloaded from ftp://ftp.ttu.edu/pub/maki.

Author's e-mail Address: bill.maki@ttu.edu

APPENDIX
Characteristics of Word Pairs Used in Rating Experiments

		JCN			FSG		
FSG	JCN	Mean	Min	Max	Mean	Min	Max
		Experiment 1					
High	0–5	1.57	0.00	3.93	0.70	0.62	0.83
	5–10	7.58	5.11	9.66	0.71	0.60	0.84
	10–15	13.40	10.11	15.00	0.70	0.62	0.84
	15–20	17.77	15.72	20.00	0.72	0.60	0.96
	> 20	21.95	20.08	26.49	0.69	0.60	0.84
Medium	0–5	1.83	0.00	4.17	0.41	0.35	0.50
	5–10	7.62	5.50	9.84	0.41	0.35	0.49
	10–15	13.07	11.44	14.68	0.41	0.35	0.49
	15–20	17.87	15.33	19.89	0.42	0.36	0.49
	> 20	21.23	20.02	22.92	0.42	0.35	0.50
Low	0–5	1.77	0.00	4.66	0.15	0.10	0.24
	5–10	8.06	5.33	9.90	0.14	0.11	0.23
	10–15	13.09	10.77	14.90	0.14	0.10	0.24
	15–20	17.27	15.49	19.60	0.15	0.11	0.21
	> 20	21.67	20.02	24.53	0.16	0.11	0.24
		Experiment 2					
High	0–4	1.58	0.00	3.49	0.72	0.61	0.91
	4–8	6.12	4.65	7.00	0.70	0.60	0.82
	8–12	9.58	8.53	11.66	0.69	0.60	0.84
	12–16	14.51	12.37	15.72	0.70	0.61	0.84
	16–20	17.88	16.01	20.00	0.70	0.60	0.88
Medium	0–4	1.59	0.00	3.57	0.42	0.35	0.49
	4–8	5.40	4.03	7.92	0.42	0.36	0.49
	8–12	10.01	8.16	11.93	0.42	0.37	0.49
	12–16	14.21	12.16	15.67	0.41	0.36	0.48
	16–20	18.55	16.94	19.82	0.43	0.36	0.50
Low	0–4	1.41	0.00	3.54	0.17	0.12	0.24
	4–8	5.38	4.01	7.71	0.14	0.11	0.22
	8–12	10.17	8.43	11.73	0.18	0.10	0.25
	12–16	14.63	12.23	15.89	0.15	0.11	0.22
	16–20	17.39	16.15	19.96	0.14	0.11	0.17

Note—JCN is the semantic distance measure; FSG is forward associative strength.

(Manuscript received December 18, 2003;
revision accepted for publication July 12, 2004.)