# Latent structure in measures of associative, semantic, and thematic knowledge

**WILLIAM S. MAKI AND ERIN BUCHANAN**
*Texas Tech University, Lubbock, Texas*

There has been much debate about the relation between knowledge for meaning (semantic memory) and knowledge for words in context (associative memory). Many measures of that knowledge exist, but do they all measure the same thing? In this study, scaling, clustering, and factor-analytic techniques were used to reveal the structure underlying 13 variables. Semantic similarity determined from lexicographic measures is shown to be separable from the associative strength determined from word association norms, and these semantic and associative measures are in turn separable from abstract representations derived from computational analyses of large bodies of text. The three-factor structure is at odds with traditional views of word knowledge. The expression of long-term knowledge about words and the concepts they represent may be better viewed in terms of associative, semantic, and thematic information.

The organization of knowledge about words and the concepts they represent has been the subject of experimentation and debate since the founding of modern experimental psychology (Esper, 1973). Our knowledge about words is expressed in associations between words, in semantic features that define concepts, and in the co-occurrence of words in written documents. For example, for most readers seeing the word MOUSE, CAT is the first associate that comes to mind. Far fewer readers will think RAT. The same readers know that MOUSE and RAT are semantically related in that they share features such as "rodent" and "has fur," but that a MOUSE eats cheese; at the same time, the readers know that a CAT resembles a MOUSE in that both have fur, but CAT is feline and eats mice. Moreover, the readers know that MOUSE and CAT could appear in the same short story, but that CAT and COMPUTER are unlikely to be featured together in a news article because they do not carry the same thematic information about computers as do COMPUTER and MOUSE. The present article is concerned with the relations among these three forms of knowledge and how they are expressed in our measures of them.

Words may become associated by virtue of their collocation in printed or spoken text (Spence & Owens, 1990), but how do such associations enter into the computation of semantics? On some accounts, the associates of a word provide a linguistic context that determines the word's meaning (Deese, 1965; Firth, 1968). On other accounts, semantics are represented in our cognitive architecture separately from word-to-word associations (Collins & Loftus, 1975; Fodor, 1983; Williams, 1996). Previous experimental and computational studies have not provided a clear resolution of this issue (for reviews, see Hutchison, 2003, and Lucas, 2000). Here we employ the statistical techniques of scaling, clustering, and factor analysis to reveal the structure underlying several measures of semantic and associative relations among pairs of words.

Advances in computational methods and resources in the last 15 years have enabled the creation of large databases for the study of word association and word meaning. Several measures of association and semantics have been computed from these databases. The oldest such measure is based on the free association technique pioneered by Galton (1879), in which subjects are instructed to give the first word that comes to mind in response to a given cue word. The largest associative database now contains over 72,000 associates of over 5,000 cue words (Nelson, McEvoy, & Schreiber, 2004), from which various measures of association have been derived (e.g., Nelson, Dyrdal, & Goodmon, 2005; Steyvers, Shiffrin, & Nelson, 2005). For example, the strength of the association linking MOUSE to CAT is .543, which is the probability of saying CAT in response to MOUSE.

Feature production is a technique similar to free association, but it is intended to elicit names of semantic features that define a target concept (McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008). The result for each tested concept (e.g., MOUSE) is a frequency distribution of feature names (e.g., *eats_cheese*). The similarity between two concepts is determined by measuring the degree to which the concepts share semantic features. Authors of dictionaries (lexicographers) perform a similar task when they invent definitions of terms in the dictionary. WordNet is an electronic dictionary now containing definitions of 155,327 words (WordNet 2.0; see Fellbaum, 1998). The dictionary is organized as a tree structure with successively higher nodes containing superordinate cat-

W. S. Maki, bill.maki@ttu.edu

egories for both nouns and verbs, the parts of speech considered here (e.g., CAT → feline → mammal, . . . ; PAY → give → transfer). Several measures of word-to-word similarity, some of which are described below, can be computed from WordNet (Patwardhan & Pedersen, 2003).

Yet another set of measures is based on computational analyses of large bodies of text (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997). All begin by processing the co-occurrences of pairs of words in pieces of text ("contexts") ranging from a single sentence to a full document, such as a news story. All then employ some statistical technique to determine the latent structure present in the word × context frequency matrix by reducing its dimensionality (from many thousands to several hundred dimensions). Similarity between words is then computed by comparing the words' locations in that high-dimensional space.

All of the measures mentioned above have been used to predict human performance in various situations, including episodic recall (Nelson et al., 2005; Steyvers et al., 2005), memory judgments (Maki, 2007), and semantic priming (Jones, Kintsch, & Mewhort, 2006). The question we raise here is, Do these measures all measure the same thing? The same question often arises in studies of individual differences (e.g., in personality or intelligence). The traditional approach has been to reduce many variables to a few "factors" that represent the latent structure underlying the different single variables. In the present research, we identified 13 variables that are thought to measure associations and/or semantics. A data set was created that contained pairs of words that had values for each of the 13 variables. The data set was analyzed using multidimensional scaling, hierarchical clustering, and exploratory factor analysis. All these methods are correlational techniques, but they offer different views of the relationships in the data. Thus, the use of multiple analytic techniques provides converging evidence for the latent structure underlying the variables.

## METHOD

Our data set consisted of 629 pairs of words for which values were available for each of five associative, three semantic, and five text-based measures. We included only variables appearing in or computed from published databases for which multiple measures could be computed.[1] Selection of word pairs for inclusion in the data set was constrained by the small size of semantic feature production norms. Feature overlap measures (cosines) were obtained for 350 noun–noun pairs from the McRae et al. (2005) norms. Another 196 noun–noun pairs were retrieved from the Vinson and Vigliocco (2008) norms. These two sets contained 42 pairs in common; the correlation between the two different cosines was very high ($r =$ .91), so the cosines were averaged for these 42 pairs. In addition, 125 verb–verb pairs were included from the Vinson and Vigliocco norms, bringing the total number of word pairs to 629.

### Associative Measures

Associative measures were based on the word association database consisting of over 72,000 associations compiled by Nelson et al. (2004). The forward strength (FSG) of a pair of words such as MOUSE–CAT is the probability of giving the word CAT to the word MOUSE. The backward strength (BSG) is the probability of giving the word MOUSE in response to the word CAT. These associative links

form a large network (a word association space; WAS) in which any two words are connected by a pathway. The associative distance between any two words is a function of the product of the strengths along the path connecting the words in that network; distances (WAS) were computed as the negative sum of the logs of the associative strengths along each path (Steyvers et al., 2005).[2] Two other associative variables, cue competitor strength (QSG) and target competitor strength (TSG), were included. These variables are computed by summing the strengths of the associates of the cue (target) word that were not connected to the target (cue) word either directly or by a single mediator (Nelson et al., 2005).

### Semantic Measures

One of the semantic feature databases includes 516 concrete nouns (McRae et al., 2005) and 2,526 features; in our study, 25 other concepts were omitted because responses were based on more than one sense—for example, *bat* (baseball) versus *bat* (animal). The second semantic feature database, the Vinson and Vigliocco (2008) norms, includes 456 words and 1,029 features and includes verb–verb as well as noun–noun pairs; 64 of the words were omitted because of their multiple senses. Each cell in the concept × feature matrix contains the probability $p(i, j)$ of feature *j*'s being given in response to concept *i*. The similarity between any two concepts is measured by computing the cosine between the two corresponding probability distributions. A cosine (COS) of 0 means that the two concepts are completely unrelated; a cosine of $+1$ means that the two concepts are identical.

Two other measures of semantic similarity were derived from WordNet 2.0 (Fellbaum, 1998). Semantic similarity in WordNet is highly dependent on the senses of the words being compared. For example, the sense of BANK depends on whether the context is MONEY or RIVER. Computer programs have been developed with which to nominate the sense for a word, given other words forming a context. We used the WordNet–SenseRelate–WordToSet package (Michelizzi & Pedersen, 2005) to make a best guess at the sense of each word in each pair of words in the data set. For a pair A–B, the dominant sense of word A was computed given word B as the context, and the dominant sense of word B was computed given word A as the context. Then the WordNet–Similarity package (Patwardhan & Pedersen, 2003) was used to compute two measures of similarity between the sense-tagged words in each pair.[3] One measure of similarity (LSK) is based on the overlap of words in the definitions of two words (Banerjee & Pedersen, 2002). A second measure of similarity (JCN) takes into account the probabilities with which the words occur in natural language; the information content of some concept *c* is $-\log p(c)$, and the difference between the information content of two concepts measures the informational distance between them (Jiang & Conrath, 1997).

### Text-Based Measures

Measures of word-to-word similarity in all three models considered here are based on a common set of documents, the Touchstone Applied Science Associates corpus (Landauer, Foltz, & Laham, 1998). In all three models, words are represented as vectors of values defining their position in high-dimensional space, but the models differ in the computational techniques used to obtain those vectors. Latent semantic analysis (LSA; Landauer & Dumais, 1997) uses singular value decomposition to reduce the dimensions computed from the word × context frequency matrix. The BEAGLE model (bound encoding of the aggregate language environment; Jones & Mewhort, 2007) uses the mathematics of light holography to store co-occurrence information. The TOPICS model (Griffiths et al., 2007) assumes that the words in a document are generated by a set of topics that are in turn generated by the gist of the document. For our analyses, we obtained measures from each of those three computational approaches from the authors of the databases. The LSA measure was based on 300 dimensions (L300). For TOPICS, we acquired measures based on 900 and 1,700 topics (T900 and T1700). For BEAGLE, we obtained a measure based on words occurring in contexts (CTX) and a composite measure that also included information about word order (CMP).

## RESULTS AND DISCUSSION

Each measure was standardized (converted to *z* scores). The transformed matrices were then subjected to three analyses—multidimensional scaling, hierarchical clustering, and exploratory factor analysis.

The Euclidean distance between each pair of variables was computed, and the resulting matrix of dissimilarities was subjected to multidimensional scaling (SPSS for Windows, Rel. 14.0.0, 2005); the scaling was restricted to two dimensions for ease of viewing. The results are shown in Figure 1 for four data sets; the combined set of 629 pairs, the 125 verb–verb pairs from the Vinson and Vigliocco (2008) norms, and the two sets of noun–noun pairs from the McRae et al. (2005) and Vinson and Vigliocco norms. The striking finding in all four analyses is the separation of the text-based measures (T900, T1700, L300, CMP, CTX) from the associative and semantic measures. Also, the associative measures (FSG, BSG, WAS, QSG, TSG) tended to cluster together, as did the semantic measures (COS, LSK, JCN). This pattern seen in the combined set (Figure 1A) is most noticeable in the verb–verb pairs (Figure 1B), but is also observed in both the sets of noun–noun pairs (Figures 1C and 1D).

The dendrogram resulting from the hierarchical clustering analysis is shown in Figure 2. Clustering was per-formed on the data for all 629 pairs, using average linkage and Euclidean distances (SPSS for Windows, Rel. 14.0.0, 2005). The highest level (topmost branch) distinguishes text-based from non-text-based measures. For the non-text-based measures, the next lower branch separates the associative measures from the semantic measures JCN and LSK, but the COS measure clusters with the competitor strength measures (QSG, TSG). We speculate that the QSG and TSG measures may reflect a tendency to give semantic features as responses during free association tests (Hutchison, 2003), so that those measures appear in the clustering analysis as similar to the COS measure.

The exploratory factor analyses were performed on the data for all 629 pairs using the FACTOR program (Lorenzo-Seva & Ferrando, 2006). We followed recommendations from methodological studies (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Our purpose was to identify latent structure, so we extracted factors using the method of unweighted least squares. A parallel analysis (Horn, 1965) recommended four factors, each with an eigenvalue greater than 1.0 and exceeding the 95% confidence interval. We aimed for a simple structure solution in which each factor had multiple variable loadings and each variable loads on only one factor (Thurstone, 1947). To achieve that goal, we applied an oblique rotation (direct oblimin), which permits correlations
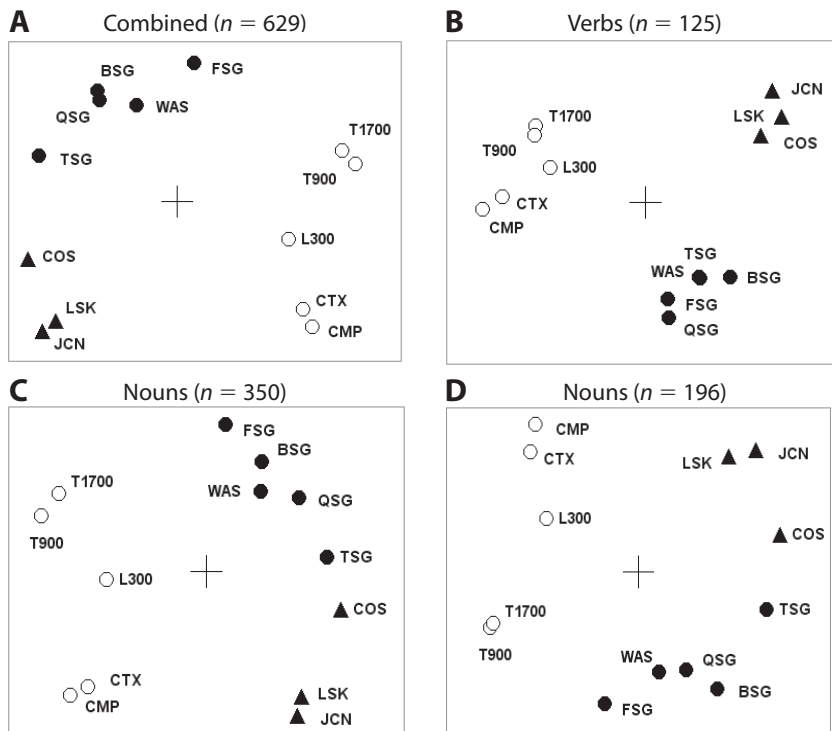


**Figure 1. Two-dimensional representations of measures of word relatedness derived from multidimensional scaling. The associative variables are forward (FSG) and backward (BSG) strengths, distance in word association space (WAS), and cue competitor strength (QSG) and target competitor strength (TSG). The semantic variables are semantic feature overlap (COS) and two measures of similarity computed from WordNet (LSK, JCN). The remaining measures are from three computational analyses of text: LSA (L300), TOPICS (T900 and T1700), and BEAGLE (CTX and CMP).**

among factors but also detects orthogonality if factor correlations are small.

The four-factor solution shown in Table 1 has the properties of simple structure. The first five (associative) variables load on the first factor, the second three (semantic) variables load on the second, and the remaining five (text-based) variables load on the remaining two. The proportion of variance explained by the four factors was .70, and the root-mean-squared residual was .04, indicating a good fit to the observed correlation matrix. Loading simplicity was .68, indicating a simple factor structure (Lorenzo-Seva, 2003). There are two main exceptions to simple structure; the COS measure loads on both of the associative and semantic factors (F1 and F2), and the LSA variable loads on both of the text-based factors (F3 and F4).

In factor-analytic studies, inclusion of multiple measured variables for each expected factor is essential (Fabrigar et al., 1999). Accordingly, we took care to include multiple measures for each of the three major types of measure. It might then be argued that the variables we selected (e.g., for the associative measures) were so highly related that the outcome was forced because of *multicolinearity* (Cohen, Cohen, West, & Aiken, 2003). One way to determine whether colinearity would force the observed factor structure is to examine tolerance values. Tolerance values are used in factor and regression analyses to measure the proportion of variance in some predictor variable that is independent of the other predictors. Tolerance values below .10 are considered indicative of colinearity (Cohen et al., 2003). In our factor analysis, none of the tolerance values are below .27, indicating that colinearity is not a likely problem with the analysis.

The colinearity objection also does not account for the split between the text-based and the non-text-based measures. LSA, TOPICS, and BEAGLE have been claimed to be models of word meaning, and thus those measures should have clustered with the other semantic measures. The separation shown in Figures 1 and 2 suggests that the text-based measures are capturing some aspect of knowledge beyond those measured by associations and defining features.

Many cue words and many response words were repeated in the 629 pairs, so one might ask whether our results are influenced by these repetitions. For each cue word with multiple responses, one response was chosen at random. Then, for each response word with multiple cues, one cue word was chosen at random. Finally, the list of pairs was randomly ordered, and pairs were sequentially selected subject to the constraint that pairs containing previously used words were rejected (thereby eliminating duplications across cue and response words). This filtering left 147 pairs with unique cues and unique responses. The reduced data set was scaled and clustered. The resulting displays exhibited three main clusters that were nearly identical to those shown in Figures 1 and 2.

The three major clusters shown in Figure 1 pose a challenge for current models of lexical processing. The split between associative and semantic measures is anticipated by models in which semantic features are stored separately from lexical representations (see, e.g., Collins & Loftus, 1975; Masson, 1995; Stolz & Besner, 1999). But such models have no obvious way of predicting the split between text-based and non-text-based measures shown in Figures 1 and 2. What are text-based measures measuring (other than, or in addition to, associative and semantic relationships)? We suggest that traditional views of separate associative and semantic representations may need to be expanded to include thematic knowledge—that is, knowledge of the gists of the stories that those words might suggest (Griffiths et al., 2007).

As we have done, our results can be interpreted within a traditional structural account that specifies separate storage of associative and semantic information. But we admit that the results reported here are open to another interpretation. Association norms, lists of defining features, and printed documents are all products of human verbal behavior, but their authors have different goals and methods. The differences in these three activities and their products may effect the latent structure revealed by our analyses. Responding to a single word is different from creating a definition, which is in turn different from writing a short story, news article, or scientific report. Such documents convey broader, more complex ideas than are present in
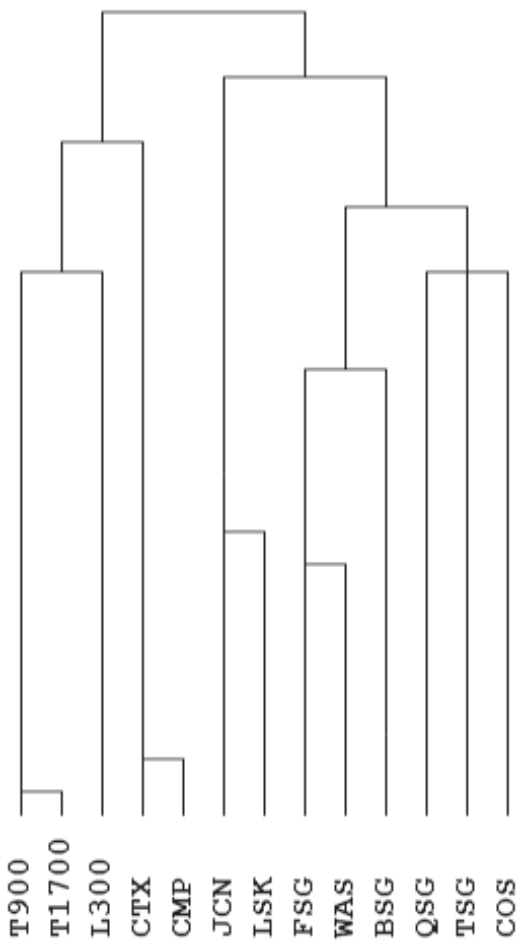


**Figure 2. Dendrogram from hierarchical clustering analysis. Variable abbreviations are the same as those in Figure 1.**

**Table 1**
**Results of Exploratory Factor Analysis Using**
**Unweighted Least Squares and an Oblique Rotation**

| Variable | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| FSG | **0.625** | −0.068 | 0.006 | 0.148 |
| BSG | **0.627** | 0.031 | 0.058 | −0.086 |
| WAS | **0.502** | 0.178 | −0.048 | 0.098 |
| QSG | **0.547** | 0.234 | −0.050 | −0.054 |
| TSG | **0.987** | −0.042 | 0.009 | −0.006 |
| COS | **0.321** | **0.389** | −0.024 | −0.023 |
| JCN | −0.056 | **0.906** | 0.004 | −0.004 |
| LSK | 0.090 | **0.659** | 0.022 | 0.018 |
| LSA | 0.219 | 0.103 | **0.301** | **0.306** |
| T900 | 0.007 | 0.007 | **1.000** | −0.006 |
| T1700 | −0.025 | −0.013 | **0.811** | −0.002 |
| CTX | −0.047 | 0.009 | −0.006 | **0.938** |
| CMP | 0.043 | −0.017 | 0.004 | **0.895** |

Note—The first five variables are from association norms, the next three variables are from feature production norms and dictionary definitions, and the last five variables are based on analyses of text. Variables (rows) and factors (columns) were organized to make factor interpretation apparent. Loadings higher than 0.3 are marked in bold.

lists of associates or defining features elicited by single words. On this account, there may be a single representation that is processed differently depending on the momentary demands of the task—guessing at associates, defining a concept, or writing a story. This ambiguity in the interpretation of our results reduces to the long-standing distinction between storage and retrieval processes. The three clusters evident in Figure 1 might arise from differently stored representations or from different retrieval operations. However this matter is resolved, our results suggest that associative, semantic, and text-based measures tap different aspects of our linguistic knowledge.

**REFERENCES**

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, February 17–23, Mexico City.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

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

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

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, **4**, 272-299.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press. Available at wordnet.princeton.edu.

Firth, J. R. (1968). A synopsis of linguistic theory 1930–1955. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth 1952–1959* (pp. 137-167). Bloomington: Indiana University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Galton, F. (1879). Psychometric experiments. *Brain*, **2**, 149-162.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, **114**, 211-244.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, **30**, 179-185.

Hutchison, K. A. (2003). Is semantic priming due to association strength or featural overlap? A *micro*analytic review. *Psychonomic Bulletin & Review*, **10**, 785-813.

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

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory & Language*, **55**, 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, **114**, 1-37.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika*, **68**, 49-60.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, **38**, 88-91.

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.

Maki, W. S. (2007). Judgments of associative memory. *Cognitive Psychology*, **54**, 319-353.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 3-23.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**, 547-559.

Michelizzi, J., & Pedersen, T. (2005). *WordNet–SenseRelate–WordToSet-0.02*. Available at search.cpan.org/dist/WordNet-SenseRelate-WordToSet.

Nelson, D. L., Dyrdal, G. M., & Goodmon, L. B. (2005). What is preexisting strength? Predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review*, **12**, 711-719.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, **36**, 402-407.

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

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

*SPSS for Windows, Rel. 14.0.0.* (2005). Chicago: SPSS Inc.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237-249). Washington, DC: American Psychological Association.

Stolz, J. A., & Besner, D. (1999). On the myth of automatic semantic activation in reading. *Current Directions in Psychological Science*, **8**, 61-65.

Thurstone, L. S. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Vinson, D., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, **40**, 183-190.

Williams, J. N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology*, **8**, 113-161.

## NOTES

1. Other measures were considered but not included. We were unable to obtain a sufficient number of values for HAL (hyperspace analogue to language; Lund & Burgess, 1996) for inclusion in our data set; the HAL Web site was nonfunctional at the time of this writing. Steyvers et al. (2005) performed multidimensional scaling on their distance measure (word association space; WAS), but only half of the cue words in the Nelson et al. (2004) norms were included because of computational constraints. Including these measures would have reduced our data set and taken our sample size below that recommended for exploratory factor analyses (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

2. For measures of distance in word association space (WAS) and distance in the dictionary (JCN), the complement of each measure was used. The transformation does not change the variance but does enable interpretation of all the variables as measures of similarity.

3. Patwardhan and Pedersen (2003) labeled their WordNet similarity measures with the same three-letter codes used here. The codes are contractions of the names of the authors of the measures: LSK (Lesk; see Banerjee & Pedersen, 2002) and JCN (Jiang & Conrath, 1997).