

WINDSORS: Windsor improved norms of distance and similarity of representations of semantics

KEVIN DURDA AND LORI BUCHANAN

University of Windsor, Windsor, Ontario, Canada

Lexical co-occurrence models of semantic memory form representations of the meaning of a word on the basis of the number of times that pairs of words occur near one another in a large body of text. These models offer a distinct advantage over models that require the collection of a large number of judgments from human subjects, since the construction of the representations can be completely automated. Unfortunately, word frequency, a well-known predictor of reaction time in several cognitive tasks, has a strong effect on the co-occurrence counts in a corpus. Two words with high frequency are more likely to occur together purely by chance than are two words that occur very infrequently. In this article, we examine a modification of a successful method for constructing semantic representations from lexical co-occurrence. We show that our new method eliminates the influence of frequency, while still capturing the semantic characteristics of words.

Visual word recognition research attempts to describe the processes by which a printed word gives rise to a pronunciation and a meaning. It is therefore not surprising that the development of metrics that quantify meaning is of particular interest to the field. An increasingly common metric for operationalizing meaning derives from computational analyses of large bodies of text that model semantic characteristics of words in the form of vectors in a high-dimensional space (Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Rohde, Gonnerman, & Plaut, 2004; Shaoul & Westbury, 2006).

These so-called *semantic spaces* can have dimensions ranging from as small as 100 to several tens of thousands. Words are represented as vectors in semantic space, and the relative distance between words is considered a reflection of their similarity of meaning. The components of the word vectors are calculated on the basis of the number of times that pairs of words occur together in a large corpus of written text. One of the earliest such models, the hyper-space analogue to language (HAL; Burgess, 2000, 2001; Burgess & Livesay, 1998; Burgess & Lund, 2000; Lund & Burgess, 1996), has enjoyed great popularity but has also been subject to much criticism. In particular, HAL has been criticized for the presence of frequency effects both within the representations (Lowe, 2001) and in the distances between word vectors (Shaoul & Westbury, 2006). The presence of frequency effects is particularly awkward for semantic metrics because the tasks used to assess visual word recognition are highly sensitive to frequency and a spurious frequency effect is likely to conceal less robust effects (such as real semantic effects).

HAL constructs vectors that represent word entries in semantic memory by analyzing the different contexts in which a word appears in a large corpus of written text. The contexts are evaluated by using a fixed-sized window that scans the text. As each word is encountered, the counts of its neighboring words are increased. These counts are weighted using a linear scheme that assigns the highest weight to the word immediately adjacent to the target, with the weight decreasing as the number of intervening words increases. These weighted counts are recorded in an $n \times n$ matrix with one row and one column assigned to each unique word. Once the matrix has been constructed, the column and row information is used to create a $2n$ vector for each word. This vector is then normalized for magnitude, and the variance of each vector component across all word vectors is calculated. Only the components with the highest variance are retained, thus creating a reduced semantic space: As few as 100 or 200 dimensions are required to model several characteristics of human semantic memory. Word similarity is analyzed in terms of the distances between word vectors, with smaller distances representing higher similarity. HAL uses Euclidean distance to determine the distance between two words.

HAL has been criticized for the strong influence of frequency on both the resulting vectors and the distances between word vectors. Shaoul and Westbury (2006) found a nonlinear function of orthographic frequency that correlated highly with distance. In addition, they found that the frequency of words varies greatly between corpora, particularly among lower frequency words. This suggests that it is vital to remove the effects of frequency from se-

mantic representations in order to limit the dependence of the representations on the particular corpus from which co-occurrence counts were collected. In selecting a set of words to form the lower dimensional semantic space, HAL chooses the words corresponding to the vector components with the highest variance across word vectors. Lowe (2001) has argued that these words will also be those that occur most often, due to the nature of the distribution of words in language. If this is the case, the vectors resulting from HAL reside in a semantic space in which many axes will correspond to the most frequently occurring words, whose role in language is usually more syntactic than semantic in nature. To support Lowe's argument, in a later section, we will offer empirical evidence that the dimension reduction step in HAL is influenced by frequency.

Despite the influences of frequency above, HAL has been successful at modeling several linguistic phenomena. The word vectors produced by HAL capture category (Burgess & Lund, 2000; Lund & Burgess, 1996) and ambiguity (Burgess, 2001) and predict behavioral data (Burgess & Lund, 2000; Conley & Burgess, 2000; Lund & Burgess, 1996; Lund, Burgess, & Atchley, 1995).

This article describes a modification of the relatively successful HAL model. We call this model WINDSORS (Windsor improved norms of distance and similarity of representations of semantics). We show that our vectors, like HAL's, model many behavioral effects but, unlike HAL's, these are free from the influences of frequency.

The effects of word frequency are most evident at the extremes of the frequency spectrum. We used two mathematical techniques to eliminate the influences of these extremes: log-relative frequency ratios (Damerau, 1993) to address high-frequency influences and scaling procedures to address low-frequency influences.

Damerau (1993) used relative frequency ratios to find words that best differentiate between two corpora. This method was proposed as a means of comparing a general text with a subject-specific text to determine a dictionary of words that are specialized to the domain of the subject. Here, we use Damerau's technique, with a Good-Turing (Good, 1953) correction,¹ to compare a word's usage in text with that same word's usage in the presence of some particular word. In effect, we use the relative frequency ratio to determine a dictionary of words that are "specialized" to a target word. With this method, we obtain frequency-free measures of word co-occurrences in a large corpus of text.

METHOD

The first stage of our method consists of collecting co-occurrence counts from a large corpus of written English text. Our corpus consisted of the British National Corpus, works from Project Gutenberg, and a collection of Apple technical documentation. The final corpus contained approximately 277 million words. Only tokens from the corpus appearing in a preselected dictionary containing 61,323 entries were counted as words; all other tokens were replaced with a nonword marker to maintain the structure of the text, but these tokens were not counted. A

window containing the 10 words on either side of a target word was passed over each word of the corpus, and the number of times that each word appeared in a window around the target was recorded in a matrix with one row and one column for each word in the dictionary. Unlike HAL (Lund & Burgess, 1996), this method does not apply different weights as a function of window position. It has been shown that searching the parameter space of window position weights is incredibly computationally expensive (Shaoul & Westbury, 2007). The different weighting schemes provide little benefit, relative to the high computational costs incurred, and due to the diversity of sentence structure, it is unlikely that a single set of weights will be appropriate in all situations.

Let t be the target word. We estimate the conditional probability of finding an instance of some word w in a window centered around an instance of the target word, denoted $\Pr(w|t)$. This can be done using the maximum likelihood estimate (MLE), calculated by dividing the number of occurrences of w and t together by the number of occurrences of t independent of context. However, if w never occurs with t in the corpus, the MLE provides an estimate of zero. This will cause the log-relative frequency ratio to be undefined for w . To address this problem, we applied one of a set of statistical techniques called *Good-Turing* methods (Good, 1953) that provide estimates of the total probability of unseen events. This probability can then be divided among the unseen events on the basis of whatever structure is present in the data. To accommodate the probability of these events, the probabilities of all observed events are adjusted, with low-frequency events having a slight reduction in probability and high-frequency events being nearly unaffected. We use the method proposed by Gale (1994), called *simple Good-Turing* (SGT), since it is simpler to use than other Good-Turing methods in practice and has been shown to be more accurate than other methods of estimating the probability of unseen events.

Because we have restricted our co-occurrence counts to only the words that appear in the preselected dictionary, the unseen events will be those words from the dictionary that were not seen with t in our corpus. Let $U(t)$ be the set of words from the dictionary that did not appear with t , and let ξ be the total probability of the words in $U(t)$ as calculated by the SGT estimate. We distribute this probability among the words in $U(t)$ in proportion to their orthographic frequency. As a word is encountered more often in a corpus, it not only occurs more often with any other given word, but also occurs with a higher number of different words, causing the co-occurrence counts to increase at a slower rate than does orthographic frequency. To account for this, we use log orthographic frequency to distribute ξ to the unseen words. Let

$$F = \sum_{w \in U(t)} \log[1 + f(w)],$$

where $f(w)$ is the orthographic frequency of w . The probability of seeing an unobserved word u is estimated as

$$\Pr(u|t) = \frac{\log[1 + f(u)]}{F} \xi.$$

For observed words, SGT provides an adjusted probability estimate that approaches the MLE as the frequency of the word increases.

We now determine whether the probability of finding w near t is greater than the probability that a word selected at random from the corpus is w , denoted $\Pr(w)$. To do this, we employ the log-relative frequency ratio

$$\Lambda(w|t) = \log \frac{\Pr(w|t)}{\Pr(w)}.$$

$\Pr(w|t)$ is estimated using SGT. $\Pr(w)$ is estimated using MLE, since every word in the dictionary occurred at least once in the corpus. Positive values of $\Lambda(w|t)$ indicate that the observed probability of w near t is greater than the probability of w independent of context.

Next, we scale the log-relative frequency ratios to account for word frequency. $\Pr(w)$ will be very small for low-frequency words. This causes small increases in $\Pr(w|t)$, such as may be produced by a single co-occurrence of w and t , to result in large relative increases in $\Lambda(w|t)$. In other words, examining co-occurrences without frequency correction tends to overestimate the importance of co-occurrence between high-frequency words. When applying the log-relative frequency ratio, we introduce the problem of overestimating the importance of co-occurrence between low-frequency words. To reduce this effect, we scale the log-relative frequency ratio by the factor $\beta(w) = 4\sqrt{1 - [2\Pr(w) - 1]^2}$. This function is monotonically increasing in $\Pr(w)$ and grows very rapidly in the low end of the frequency spectrum, allowing for greater differentiation between low-frequency events. At the opposite end of the frequency range, the function grows relatively slowly, producing similar adjustments to co-occurrences of high-frequency words. The value for the vector component corresponding to w in the vector representing t is given by

$$\Lambda_\beta(w|t) = \beta(w) \log \frac{\Pr(w|t)}{\Pr(w)}.$$

We call $\Lambda_\beta(w|t)$ the *association* of w to t . This value is calculated for each word w and then sorted alphabetically by corresponding word to produce a vector representing t in the semantic space.

The length of each vector constructed in the manner described above is equal to the number of words in the dictionary. That is, each vector consists of 61,323 components. As other co-occurrence models have demonstrated, the dimension of these vectors can be reduced to a much smaller number without sacrificing information contained within the vectors (Landauer & Dumais, 1997; Li, Burgess, & Lund, 1998; Lund & Burgess, 1996). This reduces the computational resources required to analyze relationships between vectors. To reduce these vectors, we use the same method as HAL and retain only those components that correspond to the words with the highest variance across representations. However, unlike with the HAL method, our selection is not influenced by frequency. This lower dimensional vector is the semantic representation of the target word t in our model.

Similarity between words is measured by the correlation between word vectors, with higher values represent-

ing greater similarity and negative values representing very low similarity (and not opposite meanings). In some cases, the direct use of similarity will result in an awkward mapping to effects in behavioral data. For example, priming effects will be expressed by negative values. To circumvent this inconvenience, we offer the alternative measure of distance ($1 - \text{similarity}$).

Independence of Frequency

Our primary goals were to eliminate the effects of frequency within semantic representations (vectors) and between those representations (similarity/distance). With respect to the within-vector goal, an inspection of the correlation between a word's frequency and its scaled log-relative frequency [i.e., $\Lambda_\beta(w|t)$] within a given vector reveals that we were successful: The 95% confidence interval for these correlations ranged from $-.039$ to $-.037$ for a sample of 3,409 vectors. The above demonstrates an improvement over the same correlation found in HAL data (mean correlation = .92, $SD = .06$).² Note that the HAL model used for comparisons in this article was constructed with the same corpus as our WINDSORS model.

The full vectors are frequency free, but we do not use these vectors in our final measures. The measures of interest to us are those that were derived following the dimension reduction described earlier. We therefore also tested the influence of frequency within these reduced vectors. To do this, we ranked each word according to its variance across all word vectors [i.e., the variance of $\Lambda_\beta(w|t)$ across all t s for a fixed w] and selected the top 300 items. The rank did not correlate with frequency [$r(300) = -.051$, $p = .379$]. A similar analysis conducted for HAL revealed a moderate correlation [$r(300) = -.403$, $p < .001$], confirming the analysis done by Lowe (2001). The scatterplot of rank against log frequency given in Figure 1 clearly shows that frequency has less influence in the rankings of our model (right panel) than in HAL (left panel). An unpaired samples t test confirms that the 300 items from our model were lower in average frequency ($M = 556.45$, $SD = 2,313.28$) than were those from HAL ($M = 1,910.86$, $SD = 5,391.71$) [$t(598) = -3.998$, $p < .001$].

Vector Distances

The tests of the influences of frequency on a single vector above demonstrate that the vector characterizations are relatively free from frequency. The following test shows that the same is true for our measures that involved contrasts (or matching) of vectors.

We used a sample of 33,469 pairs of words and calculated the distance between each pair. Only a small correlation between distance and orthographic frequency of the target word was found [$r(33469) = -.027$, $p < .001$], indicating that the influence of frequency on distance is small (although significant with such a large sample). Figure 2 shows a scatterplot of distance against target frequency and clearly displays that there is little relationship between the two variables.

Our goal was to develop a frequency-free measure of semantic distance for use in stimulus set development, and the tests above demonstrate that this has been achieved.

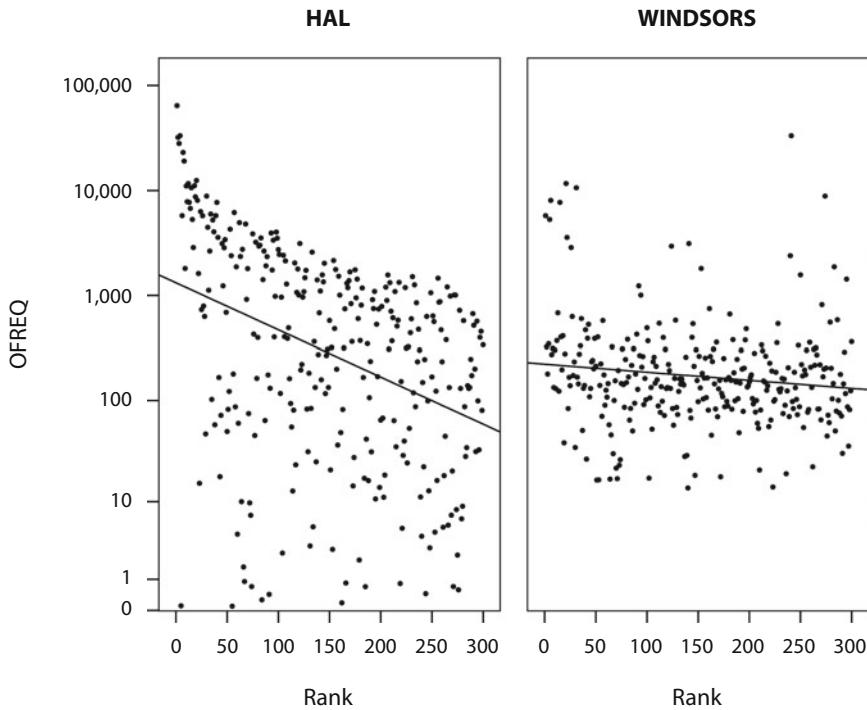


Figure 1. Top 300 words with highest component variance plotted against orthographic frequency (OFREQ) for HAL and WINDSORS models.

However, it is also important that we demonstrate that semantic content was not sacrificed in the removal of frequency. The following section describes a series of modeling experiments using data from previously published work. The experiments are those that have been successfully modeled by other high-dimensional semantic space models (i.e., HAL, LSA, and BEAGLE; Jones, Kintsch, & Mewhort, 2006).

EXPERIMENTS

Semantic-Priming Experiments

We first examined a simple semantic-priming lexical decision effect reported in Experiment 3 of McNamara and Altarriba (1988). The related word pairs given in Appendix A of McNamara and Altarriba were used for the related condition, and the unrelated condition was created by randomly rearranging the primes. To be successful, our distance measures for the related pairs would have to be smaller than the distance measures for the unrelated pairs, which they were: Mean distance for the related pairs ($M = .45$, $SD = .30$) was smaller than the mean distance in the unrelated condition ($M = .68$, $SD = .32$), and this difference was confirmed in a paired-samples t test [$t(15) = 2.21$, $p = .043$]. This simple experiment demonstrates that related words are located closer to one another than are unrelated words in the semantic space constructed by our model.

We next tested our model's sensitivity to the difference between semantic and associative relationships. Chiarello, Burgess, Richards, and Pollock (1990) com-

pared priming effects for three types of prime-target pairs: semantically similar (*deer–pony*), associatively related (*bee–honey*), and both semantically similar and associatively related (*doctor–nurse*). There was no priming in the associative condition, a robust effect in the semantically similar condition, and a still larger effect in the semantic and associative condition. We simulated their naming experiment (Experiment 2, collapsed across visual field). As in the previous simulation, we created an unrelated condition by randomly rearranging the primes. These data were analyzed in a within-subjects ANOVA with a Bonferroni correction. Our results closely matched those of Chiarello et al. The mean and standard deviation of distance for each condition are given in Table 1. There was a main effect of pair type [related or unrelated; $F(1,135) = 135.00$, $p < .001$] and prime type [associative, similar, or both; $F(2,135) = 7.00$, $p = .001$], as well as an interaction between the two [$F(2,135) = 6.89$, $p = .001$]. The mean difference between the unrelated and the related conditions with associative-only pairs was 0.1149 [$t(42) = 3.82$, $p < .001$]. In the semantic-only condition, the mean difference was 0.2064 [$t(46) = 6.34$, $p < .001$], and in the combined condition, it was 0.2753 [$t(47) = 9.76$, $p < .001$]. Within the related conditions, there was a difference between the associative-only and the associative–semantic conditions ($p < .001$) and between the semantic-only and the associative–semantic conditions ($p = .009$). The difference between the associative-only and the semantic-only conditions was not significant ($p = .097$). No differences were found in the unrelated condition (all $p > .05$). The pattern of per-

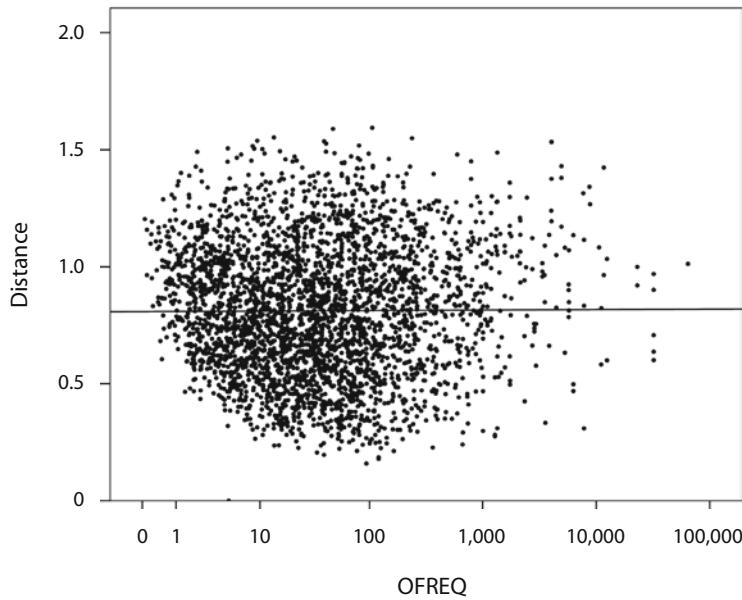


Figure 2. Plot of orthographic frequency (OFREQ) against distance for a sample of approximately 3,000 word pairs.

formance (see Figure 3) was similar to that observed in subjects by Chiarello et al.

Ferrand and New (2003) performed a similar experiment using a lexical decision task. Association and similarity were more carefully controlled between conditions, and no combined condition was included. They found a priming effect for both types of prime–target pairs, with a larger effect in the semantically similar condition. Again, our model produced a pattern of performance that mimics that of the human subjects, with a main effect of relatedness [$F(1,79) = 75.20, p < .001$]. The mean difference between the unrelated and the related conditions (i.e., the priming effect) for the associative pairs was .1551 [$t(40) = 4.48, p < .001$]. For semantically similar pairs, this difference was .2498 [$t(39) = 8.00, p < .001$], which agrees with the behavioral data (see Figure 4).

These experiments demonstrate that our vectors capture both semantic and associative relationships between words. Distance between words models priming effects in a manner similar to that for effects produced by human subjects, with weaker effects for associative pairs than for semantically similar pairs and an associative boost for pairs of words that are related by both association and semantic similarity.

Table 1
Means and Standard Deviations of Distances Between Word Pairs From Chiarello, Burgess, Richards, and Pollock (1990) Produced by WINDSORS Model

Pair Type	Prime Type	M	SD
Related	Associative	.61	.19
	Semantic	.53	.16
	Both	.42	.17
Unrelated	Associative	.72	.19
	Semantic	.73	.19
	Both	.70	.15

Word Similarity Measures

Rubenstein and Goodenough (1965) asked 51 subjects to make similarity judgments for 65 pairs of words. The subjects ranked the similarity of each pair with a number between 0 and 4, assigning a higher rank to pairs that were more similar. We calculated the similarity of the word pairs given in Table 1 of Rubenstein and Goodenough and found a moderate correlation between our similarity measure and the human norms [$r(65) = .49, r^2 = .24$]. To further evaluate the ability of our model to make these similarity judgments, we constructed a hierarchical clustering of a subset of the Rubenstein and Goodenough word pairs, using the SOTA algorithm (Herrero, Valencia, & Dopazo, 2001). As Figure 5 shows, the model paired these 24 words with high accuracy. The only misplaced pairs are *voyage*–*journey* and *coast*–*shore*.

DISCUSSION

In this article, we have presented a new method of constructing vector-based representations of the semantic content of words. We have demonstrated that the vectors produced by our model are free from the influence of orthographic frequency, both in terms of the distances between word vectors and in the construction of the vectors themselves. Our model captures aspects of human semantic memory as evidenced by the agreement between word distance and reaction time from behavioral experiments. In addition, our model has an advantage in the form of a smaller parameter space, due to the absence of a weighting function for the co-occurrence counts.

Although we have focused primarily on the presence of frequency in HAL, a preliminary analysis of the LSA model (Landauer & Dumais, 1997) reveals a similar effect of frequency on word similarity measures. We constructed

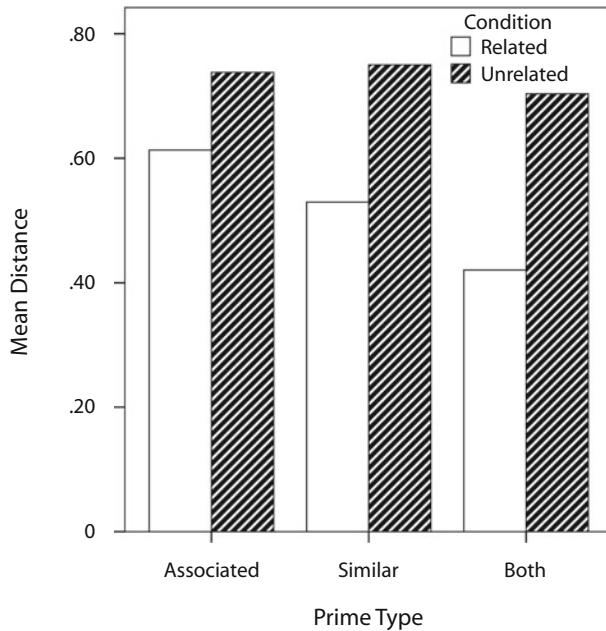


Figure 3. Average distances between word pairs from the simulation of Experiment 2 in Chiarello, Burgess, Richards, and Pollack (1990). Associatively related words produce the largest differences, with smaller differences between semantically similar pairs and smallest differences between pairs that are both semantically and associatively related. All priming effects are significant.

a list of 470 words with frequencies ranging from 1 to 1,383 per million words. A second list of 40 target words with frequencies ranging from 0.63 to 64,356 per million was created, and the cosine similarity between each of these target words and the words in the first list was retrieved from LSA.³ The correlation between these similarity measures and log-frequency of the 470 words in the first list was calculated for each of the 40 target words. These correlations ranged from .12 to .75 ($M = .39$, $SD = .17$), with larger correlations associated with the high-frequency words. We believe that the techniques used to control for frequency effects in our WINDSORS model are applicable to solving that same problem in LSA.

LSA differs from HAL in three critical ways. First, word–document co-occurrences, rather than word–word co-occurrences as in HAL, are collected from the corpus. The corpus is divided into small parcels of text, called *documents*, that are roughly the size of a paragraph (151 words, on average). The number of times that each word occurs in each document is recorded in a word–document matrix containing one row for each word and one column for each document. To reduce the effects of orthographic frequency, a step absent from HAL, the logarithm of each entry in the matrix is taken, and each entry in the row corresponding to a particular word is divided by the entropy of that word across all documents, given by $-\sum p \log p$. Finally, dimension reduction is achieved by means of singular value decomposition (SVD), rather than retaining only the components with the highest variance. The word vectors resulting from SVD are dense and low dimensional: Typically, 300 dimensions are enough to extract

the semantic relationships between words, while minimizing computational demands. Note that these dimensions do not correspond to words, as in HAL and WINDSORS, and have no direct relation to the words or documents contained in the corpus.

We propose that the scaled log-relative frequency ratio can be used in place of the frequency control step of LSA. For each document in the corpus, we can compare the frequency of each word in that document with the frequency of those same words in the full corpus to determine which words are most representative of the content of that particular document. SVD can then be performed on this transformed matrix to produce lower dimensional semantic representations. It is unlikely that using Good–Turing techniques to estimate the probability of unseen words in each document will be of any benefit, due to the extreme sparseness of the word–document matrix. The implementation of this modified LSA model is well beyond the scope of this article, and we provide this suggestion as an example that our technique may be applied to models other than HAL.

Future directions for this work include a detailed analysis of the internal structure of the representations. We hope to identify the components of the vectors that capture association and the components that capture semantic similarity. Our aim is to work toward consolidation of language-based models of semantics, such as the one discussed in this article, with feature-based models that require a large number of human judgments (McRae, Cree, Seidenberg, & McNorgan, 2005; McRae, de Sa, & Seidenberg, 1997). In addition, we will examine the quality of these representations in a connectionist model of word recognition (Harm & Seidenberg, 2004; Hinton & Shallice, 1991; Plaut, McClelland, Seidenberg, & Patterson,

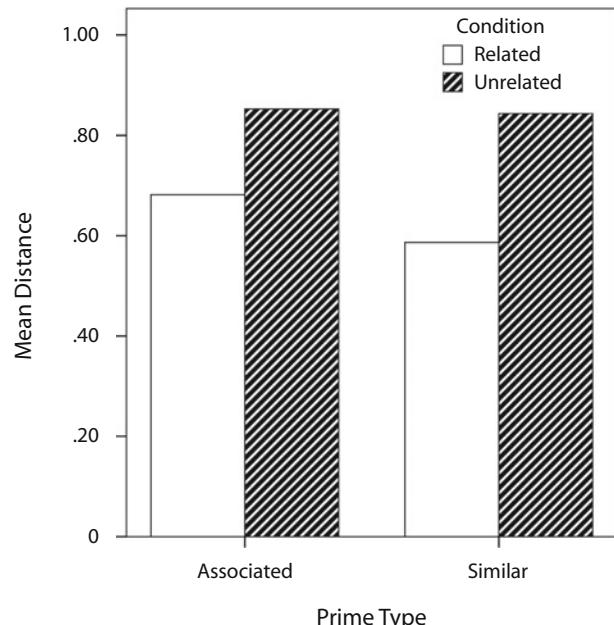


Figure 4. Results from a simulation of Ferrand and New (2003).

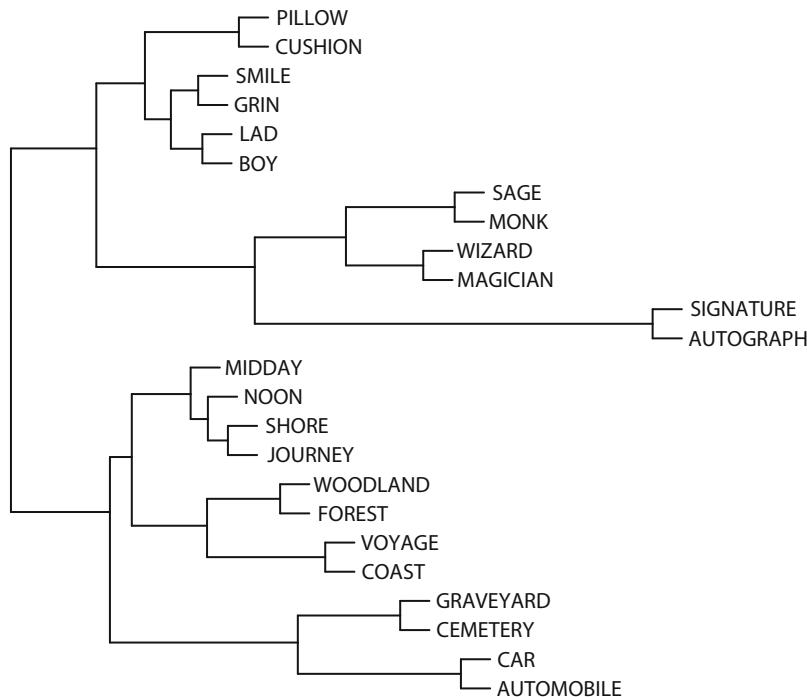


Figure 5. Hierarchical clustering of 10 pairs of semantically similar words from the Rubenstein and Goodenough (1965) data set, constructed using the SOTA algorithm.

son, 1996; Plaut & Shallice, 1993; Seidenberg & McClelland, 1989). In the past, it has been necessary to use a simplified model of semantics within a connectionist model to reduce computation time during training. Models have used randomly generated semantic representations (Rodd, Gaskell, & Marslen-Wilson, 2004), have implemented representations on the basis of only a small set of semantic features (Harm & Seidenberg, 2004; Hinton & Shallice, 1991; Plaut & Shallice, 1993), or have disregarded the semantic component completely (Perry, Ziegler, & Zorzi, 2007; Plaut et al., 1996; Seidenberg & McClelland, 1989). It is our belief that the vectors generated by the WINDSORS model will be an appropriate basis for a richer semantic component in these types of models.

AUTHOR NOTE

This research was supported by grants from NSERC to L.B. We are grateful to Richard Caron for comments on an early version of the manuscript. Correspondence concerning this article should be addressed to L. Buchanan, Department of Psychology, CHS 62, University of Windsor, Windsor, ON, N9B 3P4 Canada (e-mail: buchanan@uwindsor.ca).

REFERENCES

- BURGESS, C. (2000). Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory & Language*, **43**, 402-408.
- BURGESS, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 233-260). Washington, DC: American Psychological Association.
- BURGESS, C., & LIVESAY, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, **30**, 272-277.
- BURGESS, C., & LUND, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117-156). Mahwah, NJ: Erlbaum.
- CHIARELLO, C., BURGESS, C., RICHARDS, L., & POLLOCK, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't . . . sometimes, some places. *Brain & Language*, **38**, 75-104.
- CONLEY, P., & BURGESS, C. (2000). Age effects in a computational model of memory. *Brain & Cognition*, **43**, 104-108.
- DAMERAU, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, **29**, 433-447.
- FERRAND, L., & NEW, B. (2003). Semantic and associative priming in the mental lexicon. In P. Bonin (Ed.), *Mental lexicon: Some words to talk about words* (pp. 25-43). Hauppauge, NY: Nova.
- GALE, W. A. (1994). *Good-Turing smoothing without tears* (Statistics Research Reports from AT&T Laboratories, No. 94.5). Murray Hill, NJ: AT&T Bell Laboratories.
- GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237-264.
- HARM, M. W., & SEIDENBERG, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, **111**, 662-720.
- HERRERO, J., VALENCIA, A., & DOPAZO, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126-136.
- HINTON, G. E., & SHALICE, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74-95.
- 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.
- LI, P., BURGESS, C., & LUND, K. (1998). The acquisition of word meaning through global lexical co-occurrence. In E. V. Clark (Ed.), *Proceedings of the 30th Annual Child Language Research Forum* (pp. 167-178). Stanford, CA: Center for the Study of Language and Information.
- LOWE, W. (2001). Towards a theory of semantic space. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 576-581). Mahwah, NJ: Erlbaum.
- 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., & ATCHLEY, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 660-665). Hillsdale, NJ: Erlbaum.
- MCNAMARA, T. P., & ALTARRIBA, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory & Language*, **27**, 545-559.
- 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, Instruments, & Computers*, **37**, 547-559.
- 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.
- PERRY, C., ZIEGLER, J. C., & ZORZI, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, **114**, 273-315.
- PLAUT, D. C., McCLELLAND, J. L., SEIDENBERG, M. S., & PATTERSON, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56-115.
- PLAUT, D. C., & SHALLICE, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, **10**, 377-500.
- RODD, J. M., GASKELL, M. G., & MARSLEN-WILSON, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, **28**, 89-104.
- ROHDE, D. L. T., GONNERMAN, L. M., & PLAUT, D. C. (2004). *An improved method for deriving word meaning from lexical co-occurrence*. Unpublished manuscript. Available at dl4.mit.edu/~dr/COALS/Coals.pdf.
- RUBENSTEIN, H., & GOODENOUGH, J. B. (1965). Contextual correlates of synonymy. *Computational Linguistics*, **8**, 627-633.
- SEIDENBERG, M. S., & McCLELLAND, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523-568.
- SHAOUL, C., & WESTBURY, C. (2006). Word frequency effect in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, **38**, 190-195.
- SHAOUL, C., & WESTBURY, C. (2007, November). *Walking in space: Optimizing parameter settings in co-occurrence models of meaning*. Paper presented at the Symposium on Co-Occurrence and Lexical Organization at the 37th Annual Meeting of the Society for Computers in Psychology, Long Beach, CA.

NOTES

1. Because the logarithm is undefined at zero and many pairs of words do not occur in the corpus, we use this reliable statistical technique to estimate the probability of encountering these unobserved word pairs.

2. However, since the components of HAL vectors are computed directly from co-occurrence counts and no attempt is made to correct for frequency, this high correlation is expected.

3. LSA similarity measures were retrieved from lsa.colorado.edu, using the "general reading up to 1st year college" corpus with 300 factors. All similarities were calculated in term-term space.

(Manuscript received November 19, 2007;
revision accepted for publication March 12, 2008.)