Towards a Theory of Semantic Space

Will Lowe (wlowe02@tufts.edu)
Center for Cognitive Studies
Tufts University; MA 21015 USA

Abstract

This paper adds some theory to the growing literature of semantic space models. We motivate semantic space models from the perspective of distributional linguistics and show how an explicit mathematical formulation can provide a better understanding of existing models and suggest changes and improvements. In addition to providing a theoretical framework for current models, we consider the implications of statistical aspects of language data that have not been addressed in the psychological modeling literature. Statistical approaches to language must deal principally with count data, and this data will typically have a highly skewed frequency distribution due to Zipf's law. We consider the consequences of these facts for the construction of semantic space models, and present methods for removing frequency biases from semantic space models.

Introduction

There is a growing literature on the empirical adequacy of semantic space models across a wide range of subject domains (Burgess et al., 1998; Landauer et al., 1998; Foltz et al., 1998; McDonald and Lowe, 1998; Lowe and McDonald, 2000). However, semantic space models are typically structured and parameterized differently by each researcher. Levy and Bullinaria (2000) have explored the implications of parameter changes empirically by running multiple simulations, but there has up until now been no work that places semantic space models in an overarching theoretical framework; consequently there there are few statements of how semantic spaces *ought* to be structured in the light of their intended purpose.

In this paper we attempt to develop a theoretical framework for semantic space models by synthesizing theoretical analyses from vector space information retrieval and categorical data analysis with new basic research.

The structure of the paper is as follows. The next section brie¤y motivates semantic space models using ideas from distributional linguistics. We then review Zipf's law and its consequences the distributional character of linguistic data. The £nal section presents a formal de£nition of semantic space models and considers what effects different choices of component have on the resulting models.

Motivating Semantic Space

Firth (1968) observed that "you shall know a word by the company it keeps". If we interpret company as *lexical* company, the words that occur near to it in text or speech, then two related claims are possible. The £rst is unexceptional: we come to know about the syntactic character of a word by examining the other words that may and may not occur around it in text. Syntactic theory then postulates latent variables e.g. parts of speech and branching structure, that control the distributional properties of words and restrictions on their contexts of occurrence. The second claim is that we come to know about the *semantic* character of a word by examining the other words that may and may not occur around it in text.

The intuition for this distributional characterization of semantics is that *whatever* makes words similar or dissimilar in meaning, it must show up distributionally, in the lexical company of the word. Otherwise the supposedly semantic difference is not available to hearers and it is not easy to see how it may be learned.

If words are similar to the extent that they occur in the similar contexts then we may de£ne a statistical replacement test (Finch, 1993) which tests the meaningfulness of the result of switching one word for another in a sentence. When a corpus of meaningful sentences is available the test may be reversed (Lowe, 2000a), and under a suitable representation of lexical context, we may hold each word constant and estimate its typical surrounding context. A semantic space model is a way of representing similarity of typical context in a Euclidean space with axes determined by local word co-occurrence counts. Counting the co-occurrence of a target word with a £xed set of D other words makes it possible to position the target in a space of dimension D. A target's position with respect to other words then expresses similarity of lexical context. Since the basic notion from distributional linguistics is 'intersubstitutability in context', a semantic space model is effective to the extent it realizes this idea accurately.

Zipf's Law

The frequency of a word is (approximately) proportional to the reciprocal of its rank in a frequency list (Zipf, 1949; Mandelbrot, 1954). This is Zipf's Law. Zipf's law ensures dramatically skewed distributions for almost

all statistics applied to language; the power scaling ensures that the majority of words occur very infrequently, creating a severe sparse data problem, and that the top few most frequent words constitute the majority of all tokens. For example, the 10 most frequent word stems, or lemmas, in the 100M word British National Corpus are 'the', 'be', 'of', 'and', 'to', 'a', 'in', 'have', 'that' and 'it', constituting slightly over one quarter of all tokens in the corpus ($25974687 / 99985962 \approx 0.26$). Also the most frequent words of English are grammatical functors or closed class words (Cann, 1996), which although vital to syntax, are typically uninformative with respect to word meaning. Much of the next sections will be devoted to dealing with the distributional effects of Zipf's law.

To introduce some notation, semantic space models typically represent the distributional context of each word t in terms of a set of representative 'context' words $b_1 \ldots b_D$. t's distributional pro£le is then represented by a vector of co-occurrences $\mathbf{v}$ where $\mathbf{v}_i$ is a function of $f^W(b_i,t)$, the number of times b_i occurs in a window W words either side of t in a corpus of N words. For future reference $f(t)$ is the occurrence frequency of t in the corpus, $p(t)$ is the probability of t, often estimated by t/N, and $p^W(b_i,t)$ is the probability of seeing b_i and t together in a window of size W.

Semantic Space

A semantic space model is method of assigning each word in a language to a point in a real £nite dimensional vector space. Formally it is a quadruple $\langle \mathtt{A},\mathtt{B},\mathtt{S},\mathtt{M}\rangle$:

$\mathtt{B}$ is a set $b_{1\ldots D}$ of basis elements that determine the dimensionality D of the space and the interpretation of each dimension. $\mathtt{B}$ is often a set of words (Lund et al., 1995, e.g.) although lemmas (Lowe and McDonald, 2000), encyclopedia articles (Landauer and Dumais, 1997) and whole documents have been used.

$\mathtt{A}$ speci£es the functional form of the mapping from co-occurrence frequencies between particular basis elements and each word in the language so that each word is represented by a vector $\mathbf{v} = [\mathtt{A}(b_1,t), \mathtt{A}(b_2,t), \ldots, \mathtt{A}(b_D,t)]$. $\mathtt{A}$ may be the identity function.

$\mathtt{S}$ is a similarity measure that maps pairs of vectors onto a continuous valued quantity that represents contextual similarity.

$\mathtt{M}$, is a transformation that takes one semantic space and maps it onto another, for example by reducing its dimensionality. Various choices for these elements are possible, and lead to rather different spaces. $\mathtt{M}$ may also be an 'identity' mapping that does not change the space. In the following sections we consider the implications of different choices of $\mathtt{A},\mathtt{B},\mathtt{S}$ and $\mathtt{M}$.

$\mathtt{A}$: Lexical Association Function

Zipf's law suggest that using vectors of co-occurrence counts directly may not be a good choice when constructing a semantic space. To see why, consider two words t_1 and b with probabilities $p(t_1)$ and $p(b)$. If t_1 and b have *no* semantic relation to each other, then they will be distributionally related to one another only through their syntactic properties e.g. by the fact that they are both nouns. For simplicity we ignore any residual syntactic dependence and model their empirical frequencies $f(t_1)$ and $f(b)$ as independent binomially distributed random variables

$$\begin{aligned} f(t_1) &\sim B(p(t_1),N) \\ f(b) &\sim B(p(b),N). \end{aligned}$$

In this idealization t_1 and b are perfectly distributionally independent so $f^W(b,t_i) = WN\,p(b,t_1) = WN\,p(t_1)p(b)$ (this is just the expected co-occurrence frequency summed over each possible position in the window).

The fact that the expected co-occurrence count under independence is linear in the probability of t_1 leads to a problem in any model that sets $\mathtt{A}((b,t_i) = f^W(b,t_i)$, e.g. the Hyperspace Analogue to Language (HAL; Lund et al., 1995). Even if t_1 and t_2 are unrelated, if $p(t_1) \approx p(t_2)$ then their vectors will contain elements with similar magnitudes. This implies that any similarity measure applied to the vectors will judge them to be similar. Conversely if they are related but $p(t_1) \ll p(t_2)$ then their vectors will contain elements with widely differing magnitudes, simply due to their differing occurrence probability. Zipf's Law threatens that any difference in distributional pro£le available in $f^W(b,t_i)$ may be swamped by the effect of a difference in occurrence probability.

The upshot for models such as the HAL that use vectors of counts that are not corrected for chance is that distances will have a frequency bias. That is, proximity on semantic space will be partly due to distributional similarity, and partly due to relative frequency; the larger the difference in occurrence probability, the larger association a context element must have to affect the similarity function.

Since it is unlikely that semantic similarity depends on relative frequency, we have a theoretical reason not to use raw co-occurrence counts as a lexical association function.

Researchers in information retrieval have also noted problems with raw co-occurrence counts and use various weighting schemes to counteract them. Latent Semantic Analysis (LSA; Landauer and Dumais, 1997; Rehder et al., 1997), a semantic space model derived from information retrieval research uses an entropy-weighted function: $\mathtt{A}(b,t) \propto \log(f^W(b,t)+1)$. The logged co-occurrence count is then divided by the entropy of the distribution of b over each documents. If b is evenly distributed across documents then it is probably not informative about any particular document. In contrast if it occurs in some but not others it may be more informative about their content.

LSA's lexical association function is designed to allow arbitrarily many basis elements into the similarity calculation by weighting them appropriately. However neither logging nor dividing by entropy is guaranteed to reverse the effects of chance co-occurrence since this is never explicitly estimated.

	Target	Non-target
Context	$f^W(b,t)$	$f^W(b,\neg t)$
Non-context	$f^W(\neg b,t)$	$f^W(\neg b,\neg t)$

Table 1: Co-occurrence frequency within a window of target, context and all other words. $\neg t$ represents a word that is not t.

Lowe and McDonald (2000) used a log-odds-ratio measure to explicitly factor out chance co-occurrences. The empirical counts necessary for computing the log-odds-ratio are shown in Table 1. $\neg t$ represents any word that is not t, $\neg b$ represents a word that is not the context word b and $f^W(\neg b,t)$ is the number of times a word that is not the context word occurs among the W words surrounding t.

Computing the cell counts is straightforward because there exists a very close approximation that is a function only of $f^W(b,t)$ itself, $f(t)$, $f(b)$, W, and N:

$$\begin{aligned} f^W(b,\neg t) &= W f(b) - f^W(b,t) \\ f^W(\neg b,t) &= W f(t) - f^W(b,t) \\ f^W(\neg b,\neg t) &= WN - (f^W(b,\neg t) + f^W(\neg b,t) \\ &\quad + f^W(b,t)). \end{aligned}$$

To derive these expressions consider the limiting situation where $W = 1$ and $f(b,t)$ is the number of times the bigram $\langle b,t\rangle$ occurs. Since by de£nition $f(b) = f(b,t) + f(b,\neg t)$, then $f(b,\neg t) = f(b) - f(b,t)$, and the same reasoning applies to $f(\neg b,t)$. Similarly the number of elements in the table, $f(b) + f(\neg b)$, must be the number of bigrams in the corpus. For a large corpus this is essentially N, the number of words in the corpus. Therefore since $f(\neg b,\neg t)$ is the only cell undetermined it is obtained by subtracting the sum of the other cells from N. The W factors appear on quantities other than the co-occurrence count when the window size is more than one because only $f^W(\neg b,\neg t)$ already takes the window size into account[1].

We obtain probabilities from Table 1 by dividing each cell count by WN. Then the *odds* of seeing t rather than some other word when b is present are $p^W(b,t)/p^W(b,\neg t)$, and the odds of seeing t in the absence of b is $p^W(\neg b,t)/p^W(\neg b,\neg t)$. Therefore if the presence of b *increases* the probability of seeing t then the odds ratio (Agresti, 1990)

$$\begin{aligned} \theta(b,t) &= \frac{p^W(b,t)/p^W(b,\neg t)}{p^W(\neg b,t)/p^W(\neg b,\neg t)} \\ &= \frac{p^W(b,t)\,p^W(\neg b,\neg t)}{p^W(b,\neg t)\,p^W(\neg b,t)} \end{aligned}$$

[1]The derivation is reported elsewhere (Lowe, 2000a).

is greater than 1. When the presence of b makes no difference to the probability of seeing t then $\theta = 1$ and we can conclude that b and t are distributionally independent. Finally, if $\theta < 1$ the presence of t makes seeing b less probable.

We can estimate the odds ratio from Table 1:

$$\hat{\theta}(b,t) = \frac{f^W(b,t)\,f^W(\neg b,\neg t)}{f^W(b,\neg t)\,f^W(\neg b,t)}.$$

Where the WN factors have canceled. This measure is often logged so that then the magnitude of $\log\hat{\theta}(b,t)$ can be interpreted as a direct measure of the level of associative strength between t and b, with the effects of chance co-occurrence factored out. Positive values indicate greater than chance positive association.

Lexical Association in Lexicography

The *most* informative words for t are those that occur only in its context, e.g. t='sealed' and b='hermetically'. Instances of word pairs like this are concordances, or collocations, and are of interest to lexicographers. Consequently, the log-odds-ratio also provides a method of £nding collocations between words. Previous work in lexicography has used pointwise mutual information, log-likelihood ratios, and T-tests. Since by symmetry these alternative measures can also be lexical association functions, we review them brie¤y below.

Mutual Information The pointwise mutual information $I(b,t)$ between t and b Church and Hanks (1990) is

$$I(b,t) = \log\frac{p^W(b,t)}{Wp(b)p(t)}$$

and can be also be estimated using the frequencies in Table 1. $I(b,t)$ measures how much information an occurrence of b contains about t. If b occurs with t no more often than would be expected by chance then $p^W(b,t) = Wp(b)p(t)$ and $I(b,t) = 0$, so the mutual information measure effectively factors out random co-occurrences. However, if t and b always occur together then $p^W(b,t) = p(b)$ and $I(b,t) = \log 1/p(t)$, so the less frequent b and t are the larger their association is. In contrast, changing the marginal probabilities of t or b is equivalent to adding a constant value to rows or columns of the contingency tables above (Bishop et al., 1975). It is easy to con£rm that this change makes no difference to θ.

The G-score Dunning (1993) uses a log-likelihood ratio statistic (Agresti, 1990), which he calls the G-score, to discover collocations in text. This method compares two models of the relationship between t and b. In the £rst model (association) assumes that $p(b \mid t) \neq p(b \mid \neg t)$, whereas the second model (no association) assumes that $p(b \mid t) = p(b \mid \neg t)$. The statistic is the ratio of the maximized log-likelihoods for each model's parameters. This measure takes chance co-occurrence into account because it implicitly compares the observed co-

occurrence frequencies with the co-occurrence frequencies that would be expected by chance. For example, the expected value of the top left cell in Table 1 is $Wf(t)f(b)/N$ under (no association) but $f^W(b,t)$ under (association). Empirically using log-likelihood ratios as vector elements in a semantic space generates similar results to using log-odds-ratios. This is to be expected since both measures take chance co-occurrences into account. Alternative measures include the χ^2 statistic and Fisher's exact test. However, Dunning shows that the distributional properties of the G-score are superior under normal lexicographic conditions, and the hypergeometric probabilities required in Fisher's test are intractable to compute for contingency tables containing very large counts (Agresti, 1990). For example, $f^W(\neg b \mid \neg t)$ will typically exceed the number of words in the corpus.

Considering the lexicographic task emphasizes the 'second order' nature of semantic space measures of similarity: they re¤ect regularities across multiple '£rst order' association measures, one for each vector element. This interpretation is taken up again in discussing appropriate similarity functions below.

B : Choosing a Basis

When choosing basis elements for a semantic space there is a trade-off between choosing words that are representative of sentence content, but may not give reliable count statistics due to their low frequency, and choosing high frequency words that provide reliable statistics but appear in almost every sentence of the language. The trade-off is an instance of the bias-variance dilemma in statistical learning theory (Geman et al., 1992).

The Bias-Variance Dilemma Every statistical model is able to represent a subset of the class of possible hypotheses about data. The range of hypotheses is typically controlled by the model's structure and by a set of adjustable parameters. More ¤exible models can represent more hypotheses and are said to have less *bias*. In contrast, a very ¤exible model will require a large amount of data to determine accurate values for its parameters. When there is not enough data compared to the number of parameters, parameter estimates may be optimal for the particular data set the model was trained on, but will fail to generalize to new data. A model that 'over£ts' in this way is said to have high *variance*. Model variance can be decreased at the cost of adding bias e.g. by constraining or removing parameters. Bias can be decreased by making the model more ¤exible, at the cost of needing more data to cope with increased variance.

In a semantic space the vector elements, $\mathtt{A}(b,t)$ are parameters that estimate the amount of association between b and t on the basis of observed data $f^W(b,t)$. When choosing the basis elements $b_1 \ldots b_D$, we can de£ne a highly biased model by choosing only very high frequency words. Co-occurrence counts for high frequency words are very reliable because high frequency words appear in nearly all sentences. This biased model will have very low variance; each $\mathtt{A}(b,t)$ is a well-determined parameter because $f^W(b,t)$ is large enough to provide a reliable estimate of $p^W(b,t)$. However, every vector will be similar because all words in the language tend to occur with the high frequency words in the basis, irrespective of their distributional pro£le. Consequently, distances between words will be extremely similar and vectors in the biased model will fail to re¤ect important distributional differences.

Alternatively, if only low frequency content words are chosen as basis elements then vectors will be more highly informative and distances in the space will be able to re¤ect subtle distributional similarities. This model will have high variance because the co-occurrence counts needed to determine $\mathtt{A}(b,t)$ are unreliable. Variance can always be decreased by providing more data, but Zipf's law suggests a power relation between the amount of new text that would need to be found and the reduction in co-occurrence count variability.

In theory the fullest possible distributional pro£le for a word would include all words in the language, generating an infeasibly large vector. In practice this is not possible and some subset of words must be chosen.

The solution for LSA is to use as many words as possible with appropriate weighting for each vector element, and then use **M** to compress the original vectors into a smaller space with dimensions that are linear combinations of the original ones.

The Column Variance Method For HAL, elements of B are chosen by compiling a 70,000×70,000 matrix of word co-occurrences and discarding the columns of lowest variance[2]. Consistent with Zipf's law, column variance decreases sharply with the frequency of the word corresponding to the column (Lund et al., 1995). Then for each set of experimental stimuli, Burgess *et al.* compute variances over each vector element and retain only the most variant. We can refer to this as the column variance method of basis element choice.

The method is dif£cult to analyze because the basis is recomputed for each experiment, but we can show that it has a frequency bias. If b and t are unrelated then we can, again, model them as Binomially distributed. In the simple case where $W = 1$, the variance of the frequency count under independence is

$$\begin{aligned} \mathrm{Var}\, f^W(b,t) &= Np(t)p(b)(1-p(t)p(b)) \\ &= Np(t)p(b) - Np(t)^2p(b)^2. \end{aligned}$$

so the expected variance of $f^W(b,t)$ is quadratic in $p(b)$. The expected variance of the elements of a *column* of such counts is the same as the variance of the column sum i.e. the sum of the individual variances. Figure 1 shows the expected variances for a 14×14 table of co-occurrence counts for perfectly unrelated words with occurrence probabilities ranging from 0.5 to 0.0667. Even completely unrelated words will show distinct structure

[2] Co-occurrences are also weighted by distance, but this does not affect the following argument.

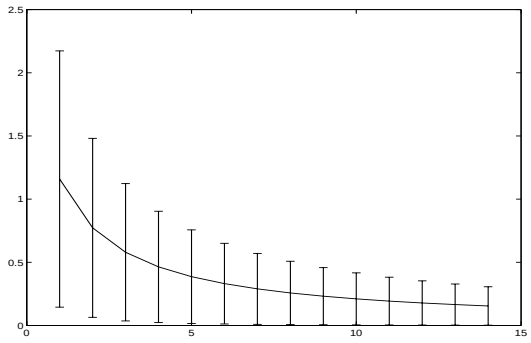

Figure 1: An example of column variance method. Expected column means based on expected co-occurrence counts between each of 14 hypothetical unrelated words. To estimate means and variances for a corpus of N words, multiply all quantities by N. Error bars represent expected column variances.

in their column variances, but this is entirely due to their baseline frequencies.

There are two possible causes for a high column variance. The £rst cause is simple frequency as shown in Figure 1. The second reason is that the words are in fact distributionally related. Then unexpectedly large variance can be a sign that the Binomial assumption has failed, and that two words are in fact related. However the size of the variance increase necessary is variable. In the column variance method, for a word that is distributionally related to some of the experimental materials to make it into the £nal basis set it must be strongly associated enough that its observed column variance moves it into the window of very high variance words at the upper end of the frequency table. In other words, it is not enough to be twice as variant as would be expected by chance, a word must be as many times more variant as it takes to have a variance that is absolutely high; lower frequency words have to work harder and unrelated but high frequency words will get chosen anyway.

This analysis of the column variance method predicts that, in the absence of strong association, the variance of a column corresponding to some candidate element will correlate strongly with that element's frequency.

This was tested by taking candidate lemmas of frequency rank 100 to 600 in the BNC, and experimental stimuli from McKoon and Ratcliff's graded priming study (see Lowe and McDonald, 2000). The analysis predicts that the levels of genuine association (corrected for frequency) between these candidates and the experimental stimuli will be be low because the words are so frequent that they provide little information about context. In fact for this data log-odds-ratios are mildly *negatively* correlated with column variance r=-.317 p<.001. In contrast candidate frequencies strongly positively correlated with column variance for co-occurrence counts, r=.8553 p<.001.

S : Similarity Measure

Two popular similarity measures are Euclidean distance and the cosine. For two vectors $\mathbf{v}$ and $\mathbf{w}$ in a D-dimensional basis, the squared Euclidean distance $\| \mathbf{v} - \mathbf{w} \|^2$ is simply related to the cosine $\rho_{\mathbf{vw}}$ of the angle between them:

$$\begin{aligned} \| \mathbf{v} - \mathbf{w} \|^2 &= \sum_{i=1}^{D} (v_i - w_i)^2 \\ &= \| \mathbf{v} \|^2 + \| \mathbf{w} \|^2 - 2 \frac{\mathbf{vw}}{\| \mathbf{v} \| \| \mathbf{w} \|} \\ &= \| \mathbf{v} \|^2 + \| \mathbf{w} \|^2 - 2 \rho_{\mathbf{vw}} \end{aligned}$$

where $\| \mathbf{w} \|^2 = \sum_i^D w_i^2$ is a squared vector length. From this equation it can be seen that $\| \mathbf{v} - \mathbf{w} \|^2 \propto \rho_{\mathbf{vw}}$ only when $\mathbf{v}$ and $\mathbf{w}$ are standardized in length. When $\mathbf{A}(b,t) = f^W(b,t)$ then vector element may have widely differing lengths depending on $p(b)$ and $p(t)$.

One advantage of the cosine is that it ranges between -1 and 1, and so removes any arbitrary scaling induced by the range of A and the number of elements in B. When A is simple co-occurrence the cosine is also less sensitive than Euclidean distance to extreme values induced by widely differing basis element frequencies, although a good choice of A should avoid this problem.

The interpretation of similarity as a 'second order' regularity can motivate yet another plausible similarity measure. We may take the correlation coef£cient (Pearson's r) as a measure of how well the elements of each word's vector match. The only difference between this and the cosine measure is that the mean of each vector is included in the similarity measure. This will not only offset the effect of different vector element magnitudes, but also place all calculations in a regular statistical framework. The statistical implications of taking correlation coef£cients over log-odds-ratios remain to be worked out. In addition, all the measures described here will bene£t from a characterization of their properties in small samples. This is future work.

M : Model

A semantic space is fully functional when a B, A and S have been speci£ed. However, it is possible to build a more structured mathematical or statistical model. In LSA the model consists of a projecting vectors into a linear subspace of B using singular value decomposition. This is equivalent to selecting the k orthogonal axes that account for most variance of words in semantic space. Each word is then projected into the the subspace, and point is then 're-in¤ated' back into the full dimensionality and cosine measures applied. Cosines can be taken in the linear subspace without subsequent re-in¤ation as suggested by Berry et al. (1995).

The theoretically important point about LSA's dimensionality reduction is that it is a simple instance of inferring latent structure in distributional data. Parts of speech, and grammatical structures are also examples of

latent structure in the sense that they are in-principle unobservable aspects of words that re¤ect their distributional properties. One important direction for semantic space research is to £nd an appropriate type of latent structure to explain the distributional regularities that are assumed to underly semantic similarity. Biologically motivated models using topographic mapping, and strictly random mappings have also been investigated (Lowe, 2000a,b).

Conclusion

In this paper we have put forward some theory for semantic space models. In addition to presenting a framework for thinking about current semantic space models we have examined the implications of various design choices, emphasized the importance of avoiding frequency biases, and presented methods for doing so. We have also connected semantic space theory to lexicographic methods and to standard problems of bias and variance discussed in the statistical literature.

Acknowledgments

Thanks to Daniel Dennett at the Center for Cognitive Studies at Tufts, where much of the work reported here was done, to the Center for Basic Research in the Social Sciences at Harvard for support, and to three anonymous referees for helpful comments.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4).

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25).

Cann, R. (1996). Categories, labels and types: Functional versus lexical. Edinburgh Occasional Papers in Linguistics EOPL-96–3, University of Edinburgh.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, (16).

Dunning, T. (1993). Accurate methods for the statistics for surprise and coincidence. *Computational Linguistics*, (19).

Finch, S. (1993). *Finding Structure in Language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Firth, J. R. (1968). A synopsis of linguistic theory. In Palmer, F. R., editor, *Selected Papers of J. R. Firth: 1952-1959*. Longman.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, (25).

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1).

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

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, (25).

Levy, J. and Bullinaria, J. (2000). Learning lexical properties from word usage patterns. In *Proceedings of the 7th Neural Computation and Psychology Workshop*. Springer Verlag.

Lowe, W. (2000a). *Topographic Maps of Semantic Space*. PhD thesis, Institute of Adaptive and Neural Computation, University of Edinburgh.

Lowe, W. (2000b). What is the dimensionality of human semantic space? In *Proceedings of the 7th Neural Computation and Psychology Workshop*. Springer Verlag.

Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, New Jersey. Lawrence Erlbaum Associates.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, (10).

McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, New Jersey. Lawrence Erlbaum Associates.

McKoon, G. and Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (18).

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., and Kintsch, W. (1997). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, (25).

Zipf, G. K. (1949). *Human Behavior and the Principal of Least Effort*. Addison Wesley.