A Comparison of Statistical Models for the Extraction of Lexical Information from Text Corpora

Simon Dennis (Simon.Dennis@colorado.edu)
Institute of Cognitive Science, University of Colorado
Boulder, Co 80301 USA

Abstract

The Syntagmatic Paradigmatic model (SP; Dennis & Harrington 2001, Dennis submitted) and the Pooled Adjacent Context model (PAC; Redington, Chater & Finch 1998) are compared on their ability to extract syntactic, semantic and associative information from a corpus of text. On a measure of syntactic class (and subclass) information based on the WordNet lexical database (Miller 1990), the models performed similarly with a small advantage for the PAC model. On a measure of semantic structure based on the similarities produced by Latent Semantic Analysis (LSA; Landauer & Dumais 1997), the models performed equivalently with a small advantage for the SP model. On a measure of associative information based on the free association norms of Nelson, McEvoy & Schreiber (1999), the SP model shows a substantive advantage over the PAC model producing more than twice as many associates.

Introduction

In recent years a number of statistical algorithms for extracting lexical information from text corpora have been proposed (Dennis submitted; Landauer & Dumais 1997, Lund & Burgess 1996, Griffiths & Steyvers 2002, Redington, Chater & Finch 1998). While each of these methods is capable of extracting lexical information from the statistics of word use, the type of information captured by each of the methods appears to differ in character. Methods such as the Syntagmatic Paradigmatic model (SP; Dennis submitted, Dennis & Harrington 2001) and the Pooled Adjacent Context model (Redington, Chater & Finch 1998) appear to group words in a way that is more indicative of syntactic class information, while models such as Latent Semantic Analysis (LSA, Landauer & Dumais 1997) and the topics model (Griffiths & Steyvers 2002) seem to extract structure that might be described as semantic. Still other models such as Hyperspace Analog to Language (HAL, Lund & Burgess 1996) appear to capture a combination of syntactic and semantic information. Work has begun on the task of systematically comparing these models (Griffiths & Steyvers 2003), but much remains to be done to characterize the type of information each of these algorithms acquire.

In this paper, the Syntagmatic Paradigmatic model (SP; Dennis, submitted, Dennis & Harrington 2002) is compared against the Pooled Adjacent Context model (PAC; Redington, Chater & Finch 1998). Both models rely on the immediately surrounding words to act as a form of context. In the SP model, each context is used as a cue to retrieve words that appear in similar contexts within the corpus. Words that commonly fill similar contexts are said to have high substitution probabilities and are deemed to be similar. By contrast, the PAC model pools the immediate contexts of a given word into a single vector. The vector corresponding to a target word is then compared against the vectors for other words to determine similarity. So, for the SP model contexts are kept separate and similarities are pooled, whereas for the PAC model contexts are pooled and then similarities calculated.

In the following sections, we first describe the SP and PAC models in more detail and provide some examples of the words that each model considers similar. Then, we contrast quantitatively their abilities to capture syntactic, semantic and associative information.

The Syntagmatic Paradigmatic Model

The Syntagmatic Paradigmatic model (SP, Dennis submitted, Dennis & Harrington 2002) is a memory-based theory of verbal cognition. It proposes that sentence processing involves the retrieval of sentence fragments from memory and the alignment of these fragments with the sentence to be interpreted. Retrieval and alignment are achieved using a Bayesian version of String Edit Theory (SET; Sankoff & Kruskal 1983).

In order to employ SET, a matrix of edit operation probabilities is induced using a version of the Expectation Maximization algorithm. These edit operation probabilities can be thought of as the lexical memory of the system, and the substitution probabilities (i.e. the probability that one word can substitute for another) can be thought of as lexical similarities. However, the EM procedure involves taking each sentence fragment from a corpus and comparing it against every other sentence fragment. Unfortunately, such a procedure is computationally expensive for large corpora where there may be tens of millions of fragments to be compared against each other.

By making a few assumptions, however, it possible to construct a fast approximation to the generic procedure. The key to improving the time complexity of the algorithm is to divide the sentence fragments into equivalence classes such that each fragment need only be compared against those from the same equivalence class rather than the entire corpus. To do this we define a fragment as a sequence of words bounded by very high frequency words (and the end of sentences boundaries) and assign fragments with the same HF word patterns to the same equivalence class. For

instance, the sentence "THE book showed A picture OF THE author carrying A copy OF THE manuscript." would be divided into the following fragments:

1. [S] THE book showed A
2. A picture OF THE
3. OF THE author carrying A
4. A copy OF THE
5. OF THE manuscript [E]

where the very high frequency words (and end of sentence markers) are marked in capital letters. Note that the second and fourth fragments would be assigned to the same equivalence class as they contain the same pattern of HF words. As a consequence, it would be deduced that "picture" and "copy" may substitute for one another. Equivalence classes are restricted to contain fragments of the same length. So, "A picture OF THE" and "A small picture OF THE" would belong to different equivalence classes.

To calculate substitution probabilities each fragment within an equivalence class was matched against each other fragment in that class. The matching strength was the count of the number of words in position that the fragments had in common. This matching strength was then normalized against the total matching strength for all of the fragments within the equivalence class. These retrieval probabilities were then averaged across the instances of each target word (appearing in different fragments).

	Match	P(Retrieval)
A picture OF THE		
A copy OF THE	3	0.33
A description OF THE	3	0.33
A side OF THE	3	0.33
ONTO THE picture [E]		
ONTO THE copy [E]	3	0.5
ONTO THE table [E]	3	0.5

P(<picture, copy>) = (0.5+0.33)/2 = 0.415

Figure 1: Illustration of SP model calculation.

Figure 1 provides an illustration of how the SP model might calculate the similarity of the word “picture” and the word “copy”. The first instance of the word “picture” appears in the fragment “A picture OF THE”. In this same equivalence class are the fragments “A copy OF THE”, “A description OF THE” and “A side OF THE”. Each of these fragments has three words in common with the retrieval cue and so each has a matching strength of 3. As there are three fragments of equal strength the retrieval probability of each fragment is 0.33, and so the substitution probability as calculated from this fragment between “picture” and “copy” is 0.33. The second instance of the word “picture” appears in the context of the fragment “ONTO THE picture [E]”. Retrieval using this fragment results in a substitution probability of 0.5, so that the average retrieval probability is 0.415.

The algorithm was run over the TASA corpus[1] using the 200 most frequent words as fragment boundaries. The corpus contains 1.2 million words, in 38000 documents and 750000 sentences. Substitution probabilities were collected for the 4000 most frequent words (note, however, that the 200 most frequent can never enter into substitutions and so in fact the substitution matrix is restricted to the 3800 subsequent words). Tables 1 and 2 show several examples of target words and the corresponding substitution candidates with the highest probabilities.

The Pooled Adjacent Context Model

The PAC model (Redington, Chater & Finch 1998) constructs a representation of a word by accumulating frequency counts of the words that appeared in the two positions immediately before and immediately after the target word. The four position vectors created in this way are then concatenated to form the representation of the word (see Figure 2).

Example Windows of Text

found	a	picture	of	the
found	a	picture	in	her
a	pretty	picture	of	her
found	a	copy	of	a
found	a	copy	below	the
destroyed	the	copy	of	the

Corresponding Pooled Vectors

picture	2	0	1	1	2	0	2	1	0	1	2	0
copy	2	1	0	0	2	1	2	0	1	2	0	1
	found	destroyed	a	pretty	a	the	of	in	below	the	her	a
	Pos -2			Pos -1			Pos 1			Pos 2		

Figure 2: Illustration of PAC model calculation

Similarities between words are determined using Spearman’s rank correlation, a form of correlation that takes into account the ranks of the values of vector components (i.e. word by position combinations). The use of the rank correlation in this case ensures that the similarities are not dominated by variability in a small number of very high frequency words.

[1] We thank the late Stephen Ivens and Touchstone Applied Science Associates (TASA) of Brewster, New York for providing this valuable resource.

Redington et. al. (1998) restricted the words for which they took frequency counts to the 150 most frequent in their corpus. In order to maintain comparability with the SP model simulations, the frequency counts of the 200 most frequent words were included, and the similarity matrix constructed includes the next 3800 most frequent words. In addition, we calculated word representations across the same TASA corpus used in the SP model simulations and did not accumulate counts across sentence boundaries Tables 1 and 2 show several examples of the highest similarity sets.

Table 1: Similarity Examples: Syntactic

Word	Ten Most Similar Words	
	SP Model	PAC Model
band	group, kind, piece, amount, lot, set, variety, series, type, line	statement, degree, bridge, hat, clock, tribe, scene, bottle, club, discussion
bands	amounts, groups, cells, pieces, patterns, natural, kinds, waves, hundreds, society	bars, columns, vapor, bases, ions, pairs, behaviors, bottles, seas, nuclei
agree	want, believe, deal, play, try, talk, begin, feel, learn, live	depend, forget, realize, listen, survive, seek, recognize, operate, discover, worry
agreed	wanted, tried, decided, believed, learned, continued, seemed, started, refused, turned	explained, answered, recognized, supported, owned, ordered, crossed, fought, removed, suggested
below	above, behind, across, among, against, near, along, inside, toward, within	above, beyond, formed, shown, higher, east, provided, watching, paid, beginning
didn't	don't, couldn't, doesn't, am, wouldn't, wasn't, can't, felt, hadn't, shall	don't, couldn't, cannot, shall, decided, fell, walked, sat, says, wasn't
myself	himself, yourself, themselves, herself, possible, meeting, going, someone, memory, want	yourself, herself, anyone, sand, wrong, walking, meat, exactly, grass, ready

Examples of Similar Words

Tables 1 and 2 present a number of examples drawn from the similarty matrices of both the SP model and the PAC model to demonstrate the different sorts of information that are extracted by each algorithm. Each row shows a word and the ten words with the highest similarities in order of similarity. The examples in Table 1 show the sensitivity of the models to syntactic categories, while the examples in Table 2 show their sensitivity to semantic and associative information.

Both models show evidence of distinguishing singular and plural nouns, past and present tense verbs, adjectives, contractions and even self pronouns. In addition, there is also evidence that both models are capturing semantic information. For instance, the most similar words to "Australia" include many countries which would not appear as associates of Australia, but are nonetheless semantically related. Likewise the most similar words for "nine" contain many numbers that are clearly semantically related. Finally, strong associates (as measured by free association) are often present in the word sets of both models. For instance, the pairs hot-cold, west-east, below-above and afternoon-evening appear in both models. In the next three sections, we will provide a quantitative assessment of how well each of the models captures – syntactic, semantic and associative information.

The Syntactic Structure Test

As a quantitative test of the ability of the models to capture syntactic structure, syntactic categories from the WordNet database (Miller 1990) were used to determine how often pairs that are deemed to be similar by each model shared a syntactic category.

WordNet classifies each word as either a noun, a verb, an adjective, a satellite adjective or an adverb. Words from closed classes are not included and some words are assigned to more than one category. For each cue word, the most similar word, the five most similar words and the ten most similar words according to each of the models were extracted.

Figure 3 shows the percentage of the time that these extracted words shared a syntactic category with the cue word according to WordNet. Both models are performing at over 90% and as the confidence intervals suggest there is no significant difference between them. There is a small and insignificant trend for the percentage to decrease as the size of the set increases.

Care must be taken in estimating chance performance to incorporate the degree of polysemy of the similar words. To ensure an appropriate baseline the target words of the substitution matrix were permuted and the analysis was repeated. Figure 3 also shows these chance baselines. Note that because each model selects a different subset of most similar words the chance baseline can vary between the two models. The SP model chose similar words that were slightly more polysemous and so the chance baseline is

marginally higher. However, again there is no significant difference between the models at any set size.

Table 2: Similarity Examples: Associative and Semantic

Word	Ten Most Similar Words	
	SP Model	PAC Model
Australia	China, India, Europe, power, Canada, California, England, America, Africa, Mexico	Philadelphia, Brazil, Florida, Kansas, Cuba, vapor, senate, males, Pennsylvania, Athens
afternoon	morning, night, year, evening, summer, room, winter, week, early, late	minute, corner, month, effort, evening, hill, conversation, chair, image, appearance
April	November, June, July, march, October, January, pages, August, September, December	June, August, dawn, Sally, Saturday, Harry, noon, Anne, Adam, Florida
diagram	picture, map, drawing, chart, book, pictures, bank, section, page, maps	graph, membrane, illustration, peninsula, valve, plateau, cord, coil, ledger, creek
hand	head, eyes, side, hands, mind, face, arms, father, arm, mouth	car, head, paper, job, name, room, line, side, child, hands
hot	cold, warm, big, fast, hard, late, strong, fresh, deep, early	heavy, cold, warm, dry, low, dark, deep, bad, simple, blue
nine	six, four, several, five, seven, eight, ten, least, twelve, twenty	twelve, fifteen, fifty, twenty, lunch, younger, rough, thirty, dinner, aunt
neutrons	protons, electrons, waves, others, atoms, animals, services, plants, metal, rays	chapters, membrane, legislature, equator, arctic, coil, valve, Mediterranean, ledger, plateau
river	sea, ocean, mountains, road, door, city, surface, floor, room, ground	tree, road, village, book, door, community, street, area, program, gas
west	north, south, east, ground, door, next, river, western, sun, morning	east, Europe, France, church, tree, town, sea, China, table, river

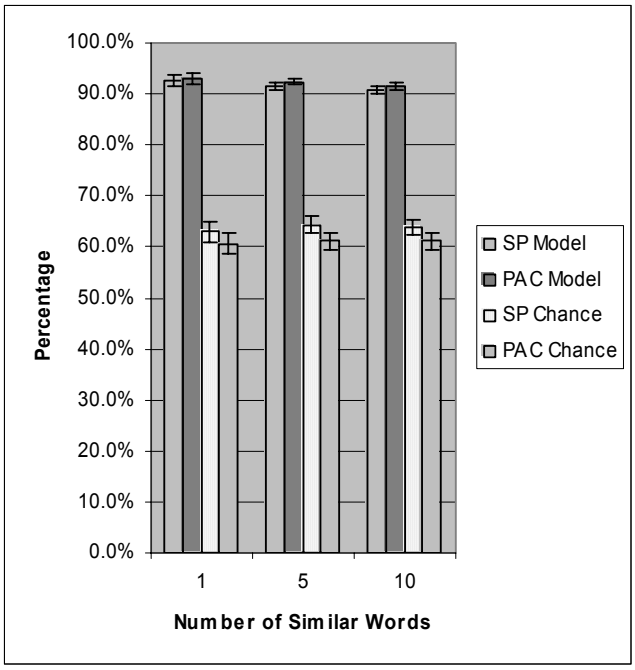

Figure 3: Percentage commonality of syntactic class for the most similar, five most similar and ten most similar words from the SP and PAC models. The bars represent 95% confidence intervals calculated using 1000 bootstrap samples (nonparametric).

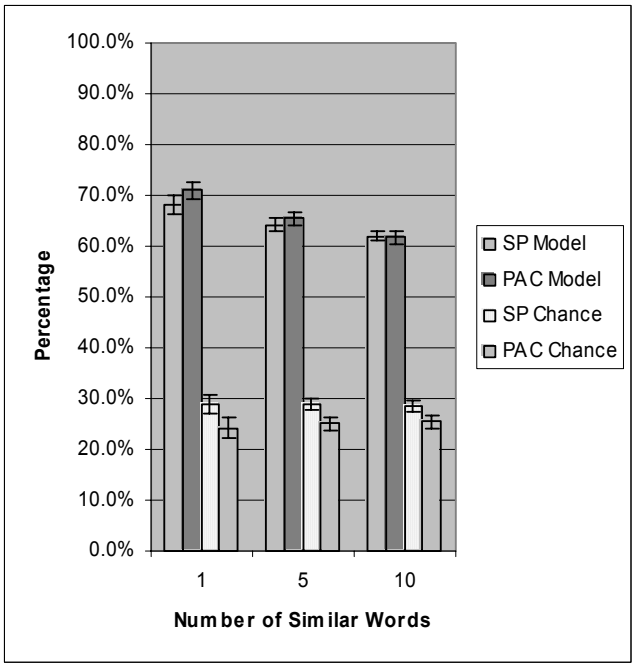

Figure 4: Percentage commonality of WordNet class for the most similar, five most similar and ten most similar words from the SP and PAC models. The bars represent 95% confidence intervals calculated using 1000 bootstrap samples (nonparametric).

In addition to syntactic categories, WordNet also reports a more fine grained classification particularly for nouns and verbs. This classification contains 45 categories and is therefore a more stringent test of the models. Figure 4 shows both the SP and PAC results when items are required to share a WordNet category. Again the models give quite similar performance, although PAC has a tendency to produce somewhat lower baseline estimates.

To summarize, the analysis suggests that both the SP model and the PAC model are capable of extracting a significant proportion of the syntactic structure, at least for high frequency words. In general, there seems to be little difference between the models with the PAC model showing a small advantage.

The Semantic Structure Test

To test the ability to capture semantic structure the most similar words produced by each model were compared for similarity based on their correspondence with the similarity cosines provided by Latent Semantic Analysis (LSA, Landauer & Dumais 1997). While data has been collected on human similarity judgments (Romney, Brewer, & Batchelder 1993) and has been used to compare models (Steyvers, Shiffrin & Nelson in progress), the available database is small in comparison to the WordNet collection or the free association norms that will be used in the next section. By contrast, LSA has been tested on a variety of tasks requiring semantic processing and provides a similarity between any two words – allowing for a comparison that is more comparable with the syntactic and associative tests provided in this paper. While it would be preferable to have a human dataset (rather than comparing against another model) current methods for collecting semantic judgments (e.g. the triads method, c.f. Romney et. al. 1993) are too intensive to be applied on a large scale.

Figure 5 shows the mean LSA cosines for the most similar, five most similar and ten most similar words produced by the SP and PAC models, respectively. Both models are performing well above chance and the analysis shows a significant but small advantage for the SP model. However, the SP model also shows an elevated mean cosine on the permuted sets of words of a similar amount (i.e. the words chosen by the SP model tend to be more similar to all other words), so there is little distinction between the models on this measure.

The Associative Structure Test

The final test compared the ability of each of the models to capture associative structure. Many early theories of association formation (Brown & Berko 1960, Ervin 1961, Ervin-Tripp 1970, McNeill 1966) proposed that associative links were formed by two basic mechanisms. Syntagmatic associations (e.g. run-fast) were thought to be acquired as a consequence of words appearing in succession in the experience of the subject. By contrast, paradigmatic associations (e.g. run-walk) were thought to occur as a consequence of experiencing words in similar sentential contexts.

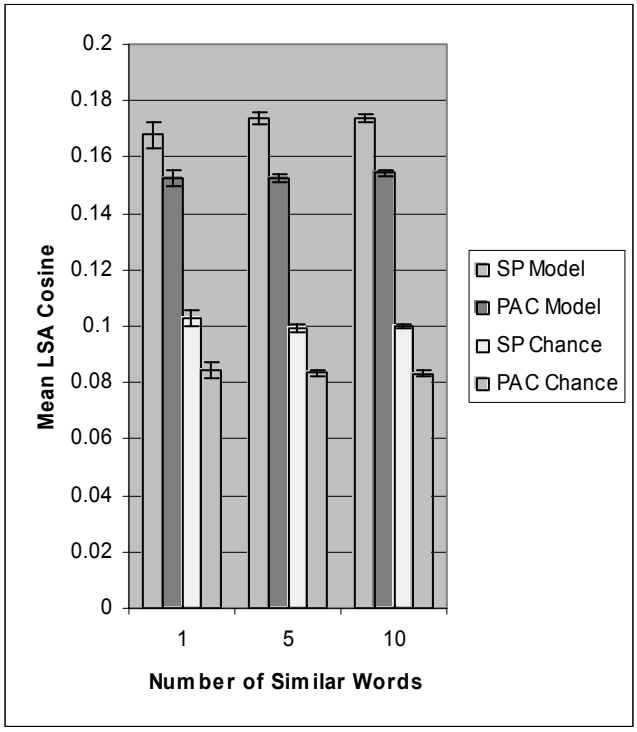

Figure 5: Mean LSA cosines for the most similar, five most similar and ten most similar words for the SP and PAC models. The bars represent 95% confidence intervals.

A systematic shift from the production of syntagmatic associates to paradigmatic associates was observed both as a consequence of development (Brown & Berko 1960, Ervin 1961) and as a function of training with nonsense syllables (McNeill 1966).

The SP model takes inspiration directly from these early models, but both the SP model and the PAC model can be considered as computational instantiations of these early ideas – particularly of the extraction of paradigmatic associates. It is of interest then, to determine to what extent they are capable of capturing free association norms.

Of the 3800 words for which statistics were calculated for each of the models, 1934 appeared in the Nelson, McEvoy and Schreiber (1999) free association norms. Figure 6 shows the count of the number of associates of these words that appeared as the most similar, in the five most similar and in the ten most words according to each of the models. The majority of the associates did not appear in the high frequency selection and so these counts are only useful as a comparison of the two models. However, the results indicate that both models are performing well above chance and that the SP model has a substantive advantage over the PAC

model, producing over twice as many associates within the 10 most similar words.

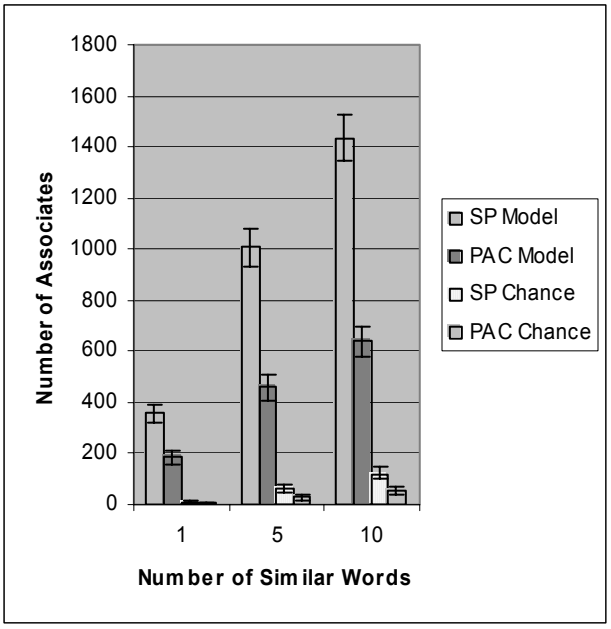

Figure 6: The number of associates among the most similar, five most similar and ten most similar words of the SP and PAC models. The bars represent 95% confidence intervals calculated from 1000 bootstrap samples (nonparametric).

Conclusion

The SP model and the PAC model share many basic assumptions. They both assume that lexical information can be induced directly from text corpora and the performance of both models on measures of syntactic structure, semantic structure and associative structure lends additional support to this conjecture (Landauer & Dumais 1997, Lund & Burgess 1996, Redington, Chater & Finch 1998).

Furthermore, both models assume that lexical similarity is determined to a large degree by the similarity of immediate sentential context (c.f. Lund & Burgess 1996). The models differ, however, in how they accumulate contextual information. In the SP model, each context in which a word appears is considered as a retrieval cue. Each instance of a word in the corpus invokes an independent memory retrieval operation and the probability of substitution is pooled across these retrievals. In the PAC model, the contexts in which a word appears are first pooled to provide a single (concatenated) vector representing the word. Similarity is then calculated by comparing these context vectors. The results suggest that while this difference has little impact on the abilities of the models to account for syntactic and semantic structure, it has a large impact on the models' ability to extract associative structure.

Acknowledgments

This research was supported by Australian Research Council grant A00106012, US National Science Foundation grant EIA-0121201 and US Department of Education grant R305G020027. I would like to thank Walter Kintsch, Tom Landauer and Jose Quesada for their helpful comments and suggestions. In addition, I would like to thank Jose Quesada for his assistance with the semantic structure analysis.

References

Brown, R. & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development, 31*, 1-14.

Dennis, S. (submitted). A memory-based theory of verbal cognition.

Dennis, S. & Harrington, M. (2001). The Syntagmatic Paradigmatic Model: An distributed instance-based model of sentence processing. The Second Workshop on Natural Language Processing and Neural Networks, 30 November, Tokyo.

Ervin, S. M. (1961). Changes with age in the verbal determinants of word association. *American Journal of Psychology*, 74, 361-372.

Ervin-Tripp, S. M. (1970). Substitution, context, and association. In L. Postman and G. Keppel (Ed.), Norms of word association. New York: Academic Press, pp. 383-467.

Griffiths, T.L., & Steyvers, M. (2002) A probabilistic approach to semantic representation. *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Griffiths, T.L., & Steyvers, M. (2003) Prediction and semantic association. *Advances in Neural Information Processing Systems 15.*

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, 105, 221-240.

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

McNeill, D. (1966) A study of word association. *Journal of Verbal Learning and Verbal Behavior, 2*, 250-262.

Miller, G. A., ed. (1990). WordNet: An On-Line Lexical Database. *International Journal of Lexicography* 3(4).

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22,* 425-469.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, 4, 28-34.

Sankoff, D. & Kruskal, J. B., eds (1983).Time warps, string edits and macromolecules: the theory and practice of sequence comparison. Addison Wesley.

Steyvers, M., Shiffrin, R.M., & Nelson, D.L. (in progress). Semantic spaces based on free association that predict memory performance.