Distributional Statistics and Thematic Role Relationships

Jon A. Willits (willits@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Sidney K. D'Mello (sdmello@memphis.edu)
Department of Computer Science, University of Memphis
Memphis, TN 38152 USA

Nicholas D. Duran (nduran@mail.psyc.memphis.edu)
Department of Psychology, University of Memphis
Memphis, TN 38152 USA

Andrew Olney (aolney@memphis.edu)
Institute for Intelligent Systems, University of Memphis
Memphis, TN 38152 USA

Abstract

Past research (McRae et al., 2005) has claimed that distributional statistics do not have enough structure to support representational relationships between thematically related nouns and verbs. We directly investigated this claim, using measures of distributional similarity. We found that several distributional statistics are sufficient not only to distinguish related from unrelated noun-verb pairs, but also more graded differences like obligatory vs. non-obligatory pairs. The consequences of these results for lexical association vs. feature-based thematic fit models are discussed, and suggestions are made for how future research might test feature-based and lexical-based versions of probabilistic constraint models of syntactic processing.

Syntactic Processing and Thematic Roles

Syntactic processing is a crucial component of language comprehension that guides the integration of linguistic elements (e.g., words, phrases, sentences) into coherent, meaningful representations. One particular area of syntactic processing that is of current interest deals with verbs and their arguments (like agents, patients, instruments, and locations). Specifically, how are these relations represented, and are simple associations between nouns and verbs sufficient to establish this relationship? McRae and colleagues (Ferretti, McRae, & Hatherall, 2001; McRae, Hare, & Tanenhaus, 2005; McRae et al, 1997; McRae et al, 2005) cite three arguments regarding the insufficiency of direct lexical association strength as the basis for thematic role comprehension: (1) lack of normative association strength between noun-verb pairs that nonetheless prime each other or facilitate reading times; (2) experimental evidence that this facilitation only occurs when nouns are in proper, role-fitting constructions (e.g. facilitation for "the cop arrested the woman" but not "the woman arrested the cop"), eliminating simple bidirectional association strength as the locus of the effect, and (3) the failure of corpus-based approaches to account for the degree of fit between a prototypical agent or patient for a specific verb (event) and the specific NP (entity or object). Rather than associative strength of nouns and verbs based on distributional information, McRae and colleagues propose that nouns and verbs have prototype representations defined in terms of semantic features, with the fit of nouns as arguments for verbs a probabilistic function of how well its features satisfy the constraints for that verb. And while McRae and colleagues assert that this prototype information is learned through both nonlinguistic conceptual experiences with objects and actions and linguistic descriptions of those objects and actions, they also claim that linguistic experience is insufficient for learning the proper roles for verbs, and that a critical part of this knowledge is in the form of conceptual, non-linguistic representations of relations between objects, actions, and events.

The main goal of this research was to determine how far simple distributional statistics can go towards capturing thematic role relationships, and to contrast the successes and failures of these statistics with McRae and colleague's feature-based thematic fit model. The success of distributional measures is relevant to assessing whether a more complex model like that of McRae et al is necessary. Conversely, any limitations of distributional statistics would also be informative, insofar as they suggest that other types of information must be learned as well.

The present research assessed the sufficiency of corpus-based distributional statistics for establishing association strengths between verbs and thematically related nouns (argument 3). If such measures can account for goodness of thematic fit, this would obviate concerns about the failures of normative association strength (argument 1). Distributional statistics may simply be more powerful predictors than association norms (see also Willits & Burgess, 2005).

McRae et al.'s strongest argument against simple associations is argument 2. However, distributional statistics need not give a simplistic, symmetric measure of association strength, and may well be able to shed light on context-sensitive activation of nouns and verbs, an issue addressed in the discussion.

Models of Distributional Statistics

Distributional statistics have been shown to be predictive of many phenomena, including grammatical categorization (Mintz, Newport, & Bever, 2001), semantic priming using the HAL model (Lund, Burgess, & Atchley, 1995), and semantic similarity judgments using the LSA model (Landauer & Dumais, 1997).

Many distributional analyses start with simple statistics (e.g., co-occurrence) and use that information to derive measures of distributional similarity. For example, Lund et al. computes a frequency-normalized co-occurrence matrix for a large set of words within some window of text (usually 8-10 words). This matrix of co-occurrences is a measure of the words' contextual usage history. Each word's row and column vectors can be thought of as a measure of the contextual usage history of that word (in the forward and backward directions). Different vectors from the matrix can be compared, giving a measure of the similarity of usage of two words. For example, the vector of co-occurrences for the words *road* and *street* will be similar because they both tend to occur with the same other words. A model of distributional similarity like that in Lund et al emphasizes the similarity of words within a narrow context, and will tend to be strongly affected by grammatical as well as semantic usage similarities.

Landauer and Dumais presented a different way of comparing two words distributional similarities. They created an frequency count matrix that is a record of the number of times a set of words occur in a set of documents. Comparing two words' vectors in their model is a measure of two words' similarity in terms of the very broad contexts (documents) in which they occur. For example, words like *stealing* and *criminal* will tend to occur together in the same documents and thus have geometrically close LSA vectors. Landauer and Dumais also used singular value decomposition (SVD), a data reduction method similar to principle components analysis, to categorize the documents in the matrix into coherent sets that tend to share the same sets of words. Prior to SVD, the similarity of words' vectors reflects a direct co-occurrence relationship between the words because it is reliant on the words co-occurring in the same documents. Performing the SVD and consolidating similar documents has the effect of finding abstract, higher-order relationships among the words. For example, in the Wikipedia corpus (used in the analyses and described below), the probability of the word *sketching* occurring in the same document as *artist* is very low (.002), and prior to SVD the similarity (cosine) of their document occurrence vectors is low (0.05). After SVD, the similarity of their vectors is 0.634. Though *artist* and *sketching* do not tend to occur together very often, they both tend to occur in documents that share many other words, and when SVD pools those documents together, the similarity of their vectors dramatically increases.

Recently, more complex statistical models have recently been proposed incorporating Bayesian techniques (Griffiths, Steyvers, & Tanenbaum, in press) and holographic memory model techniques (Jones & Mewhort, 2007). One conclusion from this work is that there are many kinds of statistics one can calculate on linguistic information, and often multiple statistics are sufficient for predicting performance for a particular linguistic phenomena. The goal of this paper is to test the simplest distributional statistics possible with regard to their sufficiency for establishing relatedness between thematically related nouns and verbs. We suggest that the simplest possible (but still relevant) distributional statistic is the likelihood of co-occurrence of the related nouns and verbs. A simple step up in complexity is the similarity of two words in terms of the other words with which they co-occur. Because we view these two types of statistics as the simplest statistics possible, these were the two used in the current study. In addition, in order to follow up on differences between sentence-based context and wider distributional contexts, and because this difference is likely to be critical to successful association of thematically related nouns and verbs, the two types of statistics will be computed on sentence-sized contexts and whole documents.

Specific Distributional Measures

All distributional measures were obtained using a corpus of text derived from the online Wikipedia (2006) encyclopedia. The entire corpus contained 1,308,712 actual articles (redirects to other articles, user discussions, and maintenance articles were not used). The corpus used consisted of 250,000 randomly chosen articles (19% of the 1.3 million). A cleaning procedure was instantiated to remove hyperlinks, special display rules, links to images, and a variety of symbols used for internal communication and text markup. Punctuations were preserved and treated as words. The cleaned corpus contained approximately 5,266,982 unique words (including many low frequency tokens like numbers, symbols, abbreviations).

For simplicity (and computational efficiency) only a small subset (10,000) of the unique tokens were used in the analyses. These words were chosen in the following two-step procedure. First, all words from a number of influential studies were included so they could be used in the analyses for this and future studies. These included all stimuli from the current investigation, as well as all items from several normative databases (McRae et al, 2004; Nelson et al, 1997). This constituted approximately 6,000 words. Second, the list was extended to 10,000 by using the most frequent words from the corpus that were not already in the list. This list was then used to create a 10,000x10,000 matrix of co-occurrences within sentences and a 10,000x250,000 matrix of occurrences within documents.

Measures of co-occurrence likelihood assess the proportion of time two words co-occur in a particular context. For this

study, the likelihood of co-occurrence within an 8-word window was computed for each noun-verb pair by taking the number of times the two words co-occurred, divided by the total number of times that word co-occurred with all 10,000 words (e.g. the vector sum). The likelihood of two words co-occurrence of two words within the same document was calculated the same way, but using the number of times the two words co-occurred in each Wikipedia document divided by the total number of times that word occurred in each document.

The distributional similarity of a word pair within an 8-word window was calculated using the HAL model. First, a 10,000-element vector of co-occurrence counts was created for each word. Next, each element was normalized using log-entropy normalization (Landauer & Dumais, 1997). Finally, the similarity of each word pair was calculated by taking the Pearson correlation of the vectors. Because the similarity of the vectors can vary significantly depending on whether the co-occurrences in the forward direction or backward direction are used, the similarity was computed both ways. This was done because it was expected that this could make a large difference on the effectiveness of the statistic with regard to thematic relatedness.

Distributional similarity within documents was calculated using the LSA model. Frequency counts were tabulated for all 10,000 words within the 250,000 documents. These frequency counts were then normalized using the log entropy frequency normalization. Each word pair's similarity was calculated by taking the Pearson correlation of the (normalized) frequency-within-document vectors. To assess the extent to which SVD is essential to extract the indirect ways in which the word pairs might be related, the correlation of the vectors was computed both before SVD (using 250,000-element normalized frequency vectors) after SVD (using the first 250 component dimensions, as is usually done with LSA).

Distributional Comparison 1

The focus of the first experiment was to determine whether verbs and nouns that are thematically related are more distributionally similar than unrelated pairs. Ferretti et al (2001) and McRae et al (2005) conducted a series of noun-verb priming experiments in order to establish that verbs and nouns that were thematically related automatically activated each other in a lexical priming experiment. Ferretti at al primed nouns with verbs that were either thematically related or unrelated in terms the nouns being good agents, patients, instruments, locations, or semantic features. They found significant priming in all cases except for locations.

McRae et al extended this work by investigating priming in the opposite direction, from nouns to thematically related verbs. McRae et al found significant priming for agents, patients, instruments, and locations. Based on their results and those of Ferretti et al, McRae et al conclude that lexical items that appropriately fit a prototypical event schema will activate each other in the appropriate contexts. These prototypical schemas are then used to aid expectancy generation during online syntactic processing.

A distributional analysis of related and unrelated words may help us evaluate the necessity of a complex model like that suggested by McRae and colleagues. Establishing what relationships exist in the input may help demonstrate the basic structure that is present in the learning environment. This structure might imply that these relationships could be learned through association and that more complicated representations might be unnecessary. Further, if and when the distributional statistics fail, this will be informative as to which additional representational structures might be necessary, and the biases in learning that might be necessary in order to bring about those representations.

Methods

Stimuli Sets of related and unrelated noun-verb pairs fitting the thematic roles for agents, patients, instruments, and locations were created using the items from Ferretti et al. and McRae et al's priming experiments. The items from both papers were pooled (with duplicates removed) resulting in 51 verb-agent pairs, 45 verb-patient pairs, 51 verb-instrument pairs, and 41 verb-location pairs. As in the priming experiments, the items were counterbalanced such that each word occurred once in the related and once in the unrelated condition.

Procedure Distributional analyses were conducted for all related and unrelated items. These analyses were conducted using the corpus and procedures described in the introduction, resulting in six dependant measures for each word pair: (1) co-occurrence likelihood within an 8-word window; (2) co-occurrence likelihood within a document; (3) forward distributional similarity within an 8-word window; (4) backward distributional similarity within an 8-word window; (5) distributional similarity within a document; (6) distributional similarity after SVD within a document. Paired *t*-tests were then computed for each dependant measure within each thematically related category. Significant *p*-values for these tests were adjusted using a Bonferroni correction to control for family-wise error rates, resulting in a critical p-value of $p = .0083$.

Results

Thematically related word pairs were more distributionally similar than unrelated word pairs for all four thematic role types (agent, patient, instrument, and location) for four of the six distributional measures. For co-occurrence likelihood, differences were statistically significant ($p < .001$) both within the 8-word window and document (see Figure 1), and for distributional similarity differences were statistically significant ($p < .001$) within a document both before and after SVD (see Figure 2).

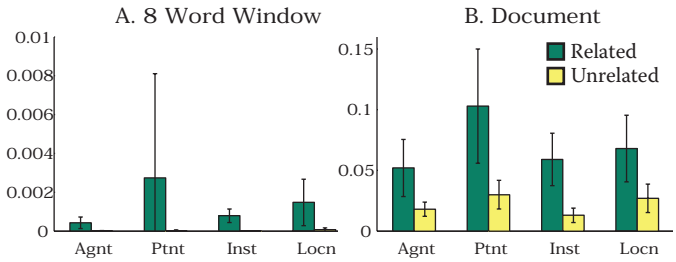

Figure 1. Co-occurrence likelihood in an 8-word window and document for thematic role relationships (Agnt = Agent, Ptnt = Patient, Inst = Instrument, and Locn = Location).

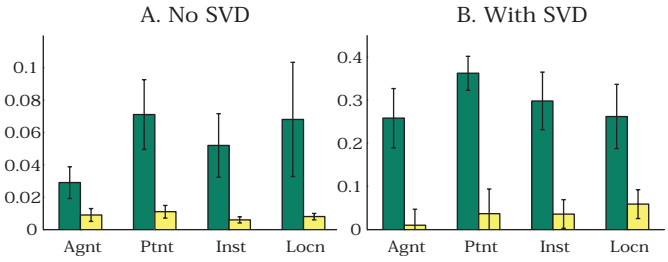

Figure 2. Distributional similarity within documents before and after SVD for thematic role relationships.

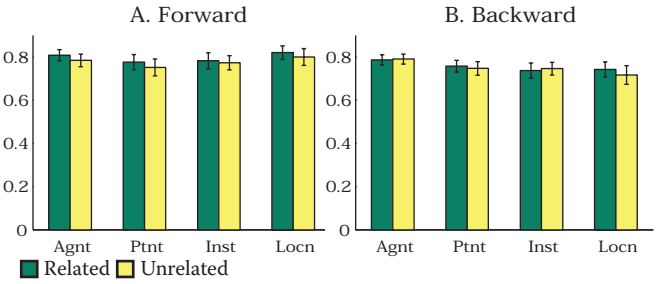

Figure 3. Forward and backward distributional similarity for 8-word window for thematic role relationships.

The distributional similarity within a sentence for related vs. unrelated pairs was not significantly different in either the forward or backward direction, neither in the forward nor backward direction (see Figure 3).

Discussion

McRae et al (2005), and historically many others (typically researchers from linguistics backgrounds or advocates of embodied cognition), have made the argument that distributional evidence is not sufficient to establish thematic role links between nouns and verbs. For at least some distributional statistics, this turns out not to be the case. Noun-verb pairs that are thematically related are more likely to co-occur with each other both within sentences and documents, and to be similar in terms of the larger contexts (documents) in which they occur. These results suggest that language learners may indeed use distributional cues to help learn the structure of event knowledge.

As argued in the introduction, the failure of the distributional statistics is often more informative than when those statistics successfully predict relationships. Beyond just pointing out that some statistics failed to be significant predictors, the pattern of distributional statistics that fail vs. those that succeed can be interesting and informative as to likely learning processes and representational structures. As such, the failure of distributional similarity within a sentence sized context (the 8-10 word window) to be a reliable predictor is very informative. The largest difference in the distributional similarity of the nouns and verbs within a sentence relative to a larger document is that the sentence-sized comparison will be much more sensitive to grammatical effects like closed class words and prepositions.

Distributional Comparison 2

Despite some researcher's skepticism, the results in Comparison 1 are not that surprising. It is not entirely shocking that words like *knife* and *cut* co-occur more often than *rag* and *cut*. A more rigorous test would compare more graded versions of thematic relatedness in addition to a binary comparison of relatedness vs. unrelatedness.

Koenig, Mueller, and Bienvenue (2003) provide an interesting set of principles for defining whether or not a noun will be a *semantic* argument of a verb (required to co-occur conceptually), or merely a semantic adjunct (allowed to conceptually co-occur, but not required). Note that this distinction is semantic, not syntactic. Under Koenig et al's view, semantic arguments of verbs will be automatically *conceptually* activated, regardless of (and perhaps in spite of) their likelihood to be present linguistically.

Koenig et al list two principles a noun must fulfill for it to be a semantic argument of a verb, only one of which will be considered in this analysis. This principle is that the noun must be obligatory to the event denoted by the verb. For example, the event behead requires an instrument, and typically a pretty specific type of instrument (like a sword). In contrast, the event kill can use an instrument, but it is not required. On this basis, Koenig et al argue that sword is an argument for behead (because it is required), but is merely an adjunct for kill (because it is optional). In a series of reading time experiments, Koenig et al's results suggest verbs which are paired with obligatory nouns aid syntactic processing better than verbs for which the same noun is only a non-obligatory adjunct. This evidence could be construed as a rule that determines precise relations between verbs and nouns. However, this relationship could also be present in distributional information, establishing a basis for learning obligatory verb-instrument pairs. And as both obligatory and non-obligatory pairs are technically thematically related (but with the obligatory pairs being in a way "more" related), it provides a good test for the ability of distributional information to account for one type of graded structure of thematic role representations.

Table 1: Distributional Statistics for Semantically Specific and Obligatory Noun-Verb Pairs

	Window Co-occurence		Document Co-occurence		→Window Dist. Sim.		←Window Dist. Sim.		Document. Dist. Sim.w/out SVD		Document Dist. Sim. with SVD	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Obligatory	.00040	.00015	.018	.007	.785	.028	.692	.030	.027	.008	.267	.044
Unobligatory	.00007	.00009	.071	.020	.767	.029	.608	.039	.018	.005	.137	.032
Mean Diff.	.00034 (p<.001)		-.052 (p<.001)		.017 (p=.274)		.084 (p=.009)		.010 (p=.076)		.132 (p<.001)	

Methods

Stimuli The stimuli are all critical verb-instrument pairs from Koenig et al (2003) Experiment 2, consisting of 24 sets of semantically obligatory/non-obligatory instruments (*sword-behead*/*kill* and *tractor-plow*/*prepare and fork-whisk*/*eat*).

Procedure The procedure is the same as in the first distributional comparison, again using a Bonferroni adjusted critical value of p = .0083.

Results

Means and standard errors for each dependant measure for each of the thematic categories are shown in Table 1. In contrast to comparison one (where related and unrelated pairs were compared) the results for obligatory and non-obligatory pairs are much more complex and potentially much more revealing. Co-occurrence likelihood within an 8-word window was significantly higher for obligatory pairs. However, co-occurrence within a document was significantly higher for *non*-obligatory pairs. Distributional similarity in a document prior to SVD was not a significantly different for obligatory vs. non-obligatory pairs (though close, p = .076). However, with an SVD, obligatory pairs were significantly more similar in terms of distributional similarity within a document. While distributional similarity within an 8-word was again not significantly different in either the forward or backward direction, the difference between means in the backward direction was 0.692 vs. 0.608 (p = .009). The size of the effect, as well as the low power (n = 24) and conservativeness of the adjustment ought to be considered when interpreting this result.

Discussion

The complexity of Comparison 2's results is intriguing. If the question is "is there a distributional statistic that can distinguish obligatory from non-obligatory verb-instrument pairs," the answer appears to be yes. Obligatory verb-instrument pairs are more likely to co-occur within a sentence-sized (8-word) context, and are more similar in terms of the larger context (document) in which they occur. They are possibly also more similar in terms of the set of words that come before them in an 8-word window. Further, it provides evidence that the criterion of whether or not a noun is obligatory for a verb is quite possibly a real construct, in so far as there are large distributional differences between obligatory and non-obligatory pairs. It also establishes that this criterion wouldn't need to be represented in rule-like form, and that it could be handled under a probabilistic constraints framework.

Again, the particular pattern of successes and failures of the different distributional statistics is of considerable interest. Particularly, the fact that arguments are more likely to co-occur within sentences, and adjuncts within documents, is very intriguing. This particular pattern of verb-instrument distribution could be very helpful for learning the entire set of obligatory and non-obligatory instruments for verb, as well as providing a basis for learning which are which.

Additionally, if distributional similarity within an 8-word window is a reliable predictor in the backward but not forward direction, this tells us something about the potential relationship of obligatory vs. optional instruments and verbs. This is especially true given the lack of both forward and backward similarity to be a reliable predictor in Comparison 1 (for related vs. unrelated words). The result raises interesting questions regarding how the 8-word window preceding verbs and obligatory instruments are similar. Do they tend to share similar function words, or similar nouns from other roles like agents or patients? Future work could follow up on this question, with interesting consequences for whether syntactic or semantic overlap are more consequential for defining the whether or not an instrument is semantically obligatory.

Finally, the fact that document similarity was not a significant predictor of a pair being obligatory until after SVD is important. This means that obligatory nouns and verbs were not actually similar in terms of the direct documents they appeared in, but were similar in terms of the *types* of documents they appeared in. This higher-order relationship implies that an important factor in whether or not a word is obligatory is their joint relationship to a large set of other words (those that define similar documents during the SVD process). Like the 8-word window results, these results could be followed up to see what other words are helped create the coherent SVD-reduced documents in which obligatory nouns and verbs were likely to co-occur, and to investigate how these words related to the particular noun-verb pairs.

General Discussion

In summary, we have provided evidence that, contrary to prior claims, some distributional statistics may be sufficient for establishing which nouns and verbs are thematically related, and even to establish more specific differences like which are obligatory and which are not. This evidence

suggests that model involving direct lexical associations could play a large part in explaining thematic-relatedness effects. However, there are two major issues to address with regard to the sufficiency of a lexical association-based model, and such a model's relation to a feature-based thematic fit model or a syntactic-rule based model.

The first question would be how a lexical association model would accommodate context sensitive activation and directional asymmetries. Ferretti et al's (2001) final experiment demonstrated that verb-noun priming only seems to occur when the related pair is being used in a contextually appropriate way (e.g. priming for *cop* in "*she was arrested by the cop*" but not in "*she arrested the cop*"). Similarly, there are some asymmetries in direction in terms of thematic role activation. McRae et al (2005) found priming from nouns to verbs for location-verb pairs, but Ferretti et al did not find priming in the verb-noun direction.

An insufficiently simple model of lexical associations would have a single association strength between the noun-verb pair (whether it was based on distributional similarity of co-occurrence likelihood). However, a more complex model incorporating several statistics could begin to produce asymmetric or context-sensitive effects. One way lexical associations could provide context sensitivity is to note that statistics are not bidirectional, and that English has many word order biases that, in combination with directional differences in probability (e.g. likelihood to co-occur before vs. after) could bootstrap knowledge about whether a noun is a likely agent or patient of a verb. Another way lexical associations could be context sensitive or asymmetric during syntactic processing is to not restrict such associations to noun-verb pairs, and to investigate the distributional patterns of the nouns and verbs with other words in the sentence. For example, one particular joint set of nouns, verbs, and particular function words may more likely for agent relationships than for patient relationships, and vice versa. Such a lexicalist model would be quite similar to lexical proposals for language and grammatical acquisition like those put forward by Bates and MacWhinney (1982).

The other major roadblock to an association-based account of thematic relatedness issue deals with the nature of the graded structure of thematic relationships. Different measures of distributional structure provide many different types of graded structure. Some are highly influenced by syntax (or function words), whereas others are more influenced by shared content words or higher-order structure. The relationship of this distributional structure and feature-based thematic models needs to be examined. It is likely that these graded structures will be highly correlated, but also distinct, and it is unclear which will be more of a match with the kind of graded effects subjects demonstrate in syntactic processing experiments. Studies investigating this question will be able to directly test distributional-based association models against feature-based thematic fit models, perhaps showing how distributional relationships and feature-based semantic models interact.

References

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Warner and L. Gleitman (Eds.), *Language acquisition*: The state of the art. New York: Cambridge University Press.

Ferretti, T. R., McRae, K, & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memoryand Language*, *44*, 516-547.

Griffiths, T. L., Steyvers, M., & Tanenbaum, J. B. (in press). Topics in semantic representation. *Psychological Review*.

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

Kintsch, W. and Bowles, A. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 17*, 249-262.

Koenig, J., Mauner, G., & Bienvenue, B. (2003). Arguments for adjuncts. *Cognition*, *89*, 67-103.

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

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *Cognitive Science Proceedings*, LEA. pg. 660-665

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language & Cognitive Processes*, *12*, 137-176.

McRae, K., Hare, M., & Tanenhaus, M. (2005). Meaning Through Syntax is insufficient to explain comprehension of sentences with reduced relative clauses: Comment on McKoon and Ratcliff (2003). *Psychological Review, 112*, 1022-1031.

McRae, K., Hare, M., Elman, J, & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33*, 1174-1184.

Willits, J. A., & Burgess, C. (2005). Semantic and associated relationships: By-products of the learning environment? Talk presented at Psychonomics Society Annual Meeting. Toronto, CA.